Advances in Survival Analysis and Optimal Scaling Methods

Willems, S.J.W.

Optimal scaling for survival analysis with ordinal data

3

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

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

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

3.1 Introduction

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

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

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

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

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

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

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

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

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

3.2 Current practice

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

3.2.1 Ordinal data in survival analysis

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

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

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

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

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

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

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

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

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

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

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

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

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

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

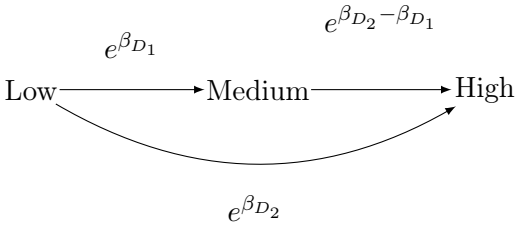
Table 3.2: *Three categories' hazard functions in dummy coding, integer coding and the optimal scaling method. Indices "D_1" and "D_2" indicate dummies 1 and 2, index "int" indicates integer coding, and index "os" indicates optimal scaling.*

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

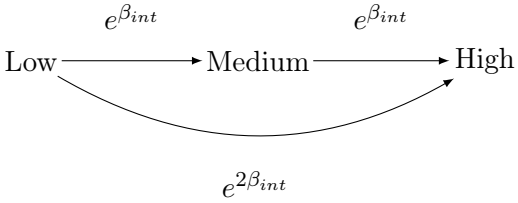
The relative risks between each category are shown in Figure 3.1a. If $\beta_{D_1}, \beta_{D_2} > 0$ or $\beta_{D_1}, \beta_{D_2} < 0$, and $|\beta_{D_2}| > |\beta_{D_1}|$ the relative risk between category levels *low* and *high* will be larger than the relative risk between levels *low* and *medium*,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

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

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



(a) *Dummy coding method.*



(b) *Integer coding method.*

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

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

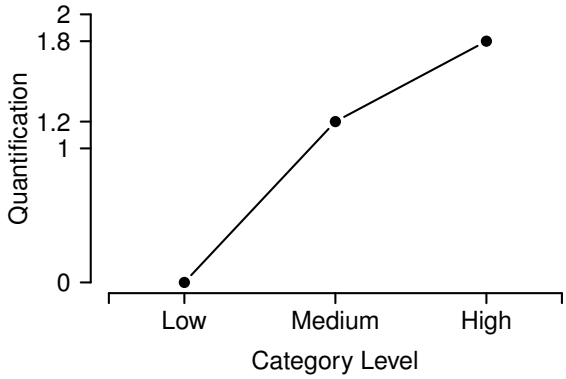


Figure 3.2: *Example of quantifications for a categorical variable with three levels.*

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

3.2.2 Optimal scaling in simple linear regression

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

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

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

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

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

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

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

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

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

The loss function is minimized by the ordinary least squares solution

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}. \tag{3.5}$$

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

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

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

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

This results in the following loss function

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

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

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

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

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

which yields

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

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

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

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

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

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

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

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

$$\check{\boldsymbol{v}}_k = \beta_k^{-1} \boldsymbol{D}_k^{-1} \mathbf{G}_k^T \boldsymbol{u}_k, \tag{3.10}$$

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

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

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

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

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

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

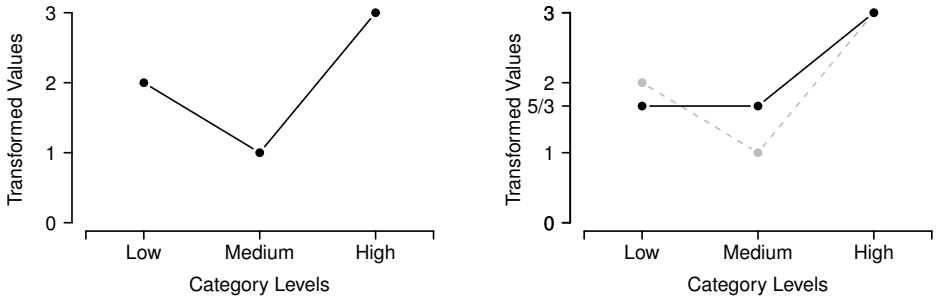
Table 3.3: *Weighted monotone regression algorithm on small data example in which the observed categories were* low*,* low*,* medium*, and* high*.*

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

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

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

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



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

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

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

Optimal scaling regression algorithm:
Step 1: Initialize $\widetilde{\beta}_k$ and $\widetilde{\mathbf{v}}_k$ for $k = 1, \ldots, p$.
Step 2: For $k = 1, \ldots, p$, do:
Step 2a: Calculate $\widetilde{\boldsymbol{u}}_k = \boldsymbol{y} - \sum_{l \neq k} \widetilde{\beta}_l \mathbf{G}_l \widetilde{\mathbf{v}}_l$.
Step 2b: Find $\check{\mathbf{v}}_k$ minimizing

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

Standardize $\check{\mathbf{v}}_k$, and denote the standardized version by $\bar{\mathbf{v}}_k$. If covariate Z_k is ordinal, apply ordinal restrictions on $\bar{\mathbf{v}}_k$, resulting in restricted quantifications $\widehat{\mathbf{v}}_k$. Set $\widetilde{\mathbf{v}}_k = \widehat{\mathbf{v}}_k$, and $\widetilde{\varphi}_k(\mathbf{Z}_{*k}) = \mathbf{G}_k \widetilde{\mathbf{v}}_k$.
Step 2c: Find $\widehat{\beta}_k$ minimizing

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

Set $\widetilde{\beta}_k = \widehat{\beta}_k$.
Step 3: Repeat step 2 until convergence of $\widetilde{\mathbf{v}}_k$ and $\widetilde{\beta}_k$.

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

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

3.3 Optimal scaling in survival analysis

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

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

3.3.1 A reweighted least squares approach to fitting the Cox model

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

proportional hazards model is defined as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

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

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

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

This likelihood can be maximized by maximizing its log,

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

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

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

where

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\widehat{\beta}_k = \left(\varphi_k(\mathbf{Z}_{*k})^T\boldsymbol{\Omega}(\boldsymbol{\eta}^*)^T\varphi_k(\mathbf{Z}_{*k})\right)^{-1}\varphi_k(\mathbf{Z}_{*k})^T\boldsymbol{\Omega}(\boldsymbol{\eta}^*)\boldsymbol{u}_k. \tag{3.31}$$

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

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

with

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

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

Optimal scaling algorithm for Cox' proportional hazards model:
Step 1: Initialize $\widetilde{\beta}_k$ and $\widetilde{\mathbf{v}}_k$ for $k = 1, \ldots, p$.
Step 2: For $k = 1, \ldots, p$, do:
Step 2a: Compute $\ell'(\widetilde{\boldsymbol{\eta}}^*)$ and $\ell''(\widetilde{\boldsymbol{\eta}}^*)$, and use these quantities to derive $\omega(\widetilde{\boldsymbol{\eta}}^*)$ and $\zeta(\widetilde{\boldsymbol{\eta}}^*)$.
Step 2b: Calculate $\boldsymbol{u}_k = \boldsymbol{\zeta}(\boldsymbol{\eta}^*) - \sum_{l \neq k} \beta_l \varphi_l(\mathbf{Z}_{*l})$.
Step 2c: Find $\check{\mathbf{v}}_k$ minimizing

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

Standardize $\check{\mathbf{v}}_k$, and denote the standardized version by $\bar{\mathbf{v}}_k$. If covariate Z_k is ordinal, apply ordinal restrictions on $\bar{\mathbf{v}}_k$, resulting in restricted quantifications $\widehat{\mathbf{v}}_k$. Set $\widetilde{\mathbf{v}}_k = \widehat{\mathbf{v}}_k$, and $\widetilde{\varphi}_k(\mathbf{Z}_{*k}) = \mathbf{G}_k \widetilde{\mathbf{v}}_k$.
Step 2d: Find $\widehat{\beta}_k$ minimizing

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

Set $\widetilde{\beta}_k = \widehat{\beta}_k$.
Step 3: Repeat step 2 until convergence criterium (3.32) is met.

3.4 Simulation study

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

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

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

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

3.4.1 Set up of simulation study

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

Generating the data

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

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

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

$$h_C = h_{0_C}. \tag{3.34}$$

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

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

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

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

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

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

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

Fitting the models

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

$$\begin{array}{rcccl}
\widehat{h}(t|Z=0) & = & \widehat{h}_{0_D}(t)\exp(0\,\widehat{\beta}_{D_1}+0\,\widehat{\beta}_{D_2}+\ldots 0\,\widehat{\beta}_{D_{C-1}}) & = & \widehat{h}_{0_D}(t)\\
\widehat{h}(t|Z=1) & = & \widehat{h}_{0_D}(t)\exp(0\,\widehat{\beta}_{D_1}+1\,\widehat{\beta}_{D_2}+\ldots 0\,\widehat{\beta}_{D_{C-1}}) & = & \widehat{h}_{0_D}(t)\exp(\widehat{\beta}_{D_1})\\
\vdots & & \vdots & & \vdots\\
\widehat{h}(t|Z=C-1) & = & \widehat{h}_{0_D}(t)\exp(0\,\widehat{\beta}_{D_1}+0\,\widehat{\beta}_{D_2}+\ldots 1\,\widehat{\beta}_{D_{C-1}}) & = & \widehat{h}_{0_D}(t)\exp(\widehat{\beta}_{D_{C-1}}).
\end{array}$$

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

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

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

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

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

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

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

Comparing performance

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

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

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

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

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

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

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

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

3.4.2 Results

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

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

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

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

Monotonically increasing model parameters

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

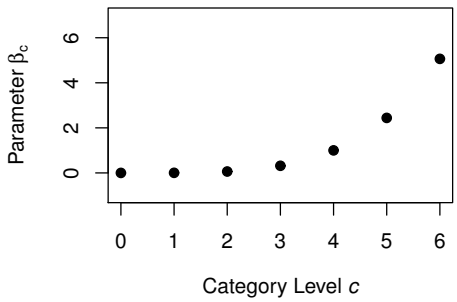


Figure 3.4: *Simulation parameters* $\beta_{Z_0}, \beta_{Z_1}, \ldots, \beta_{Z_6}$.

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

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

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

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

Increasing sample size

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

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

Increasing censoring percentage

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

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

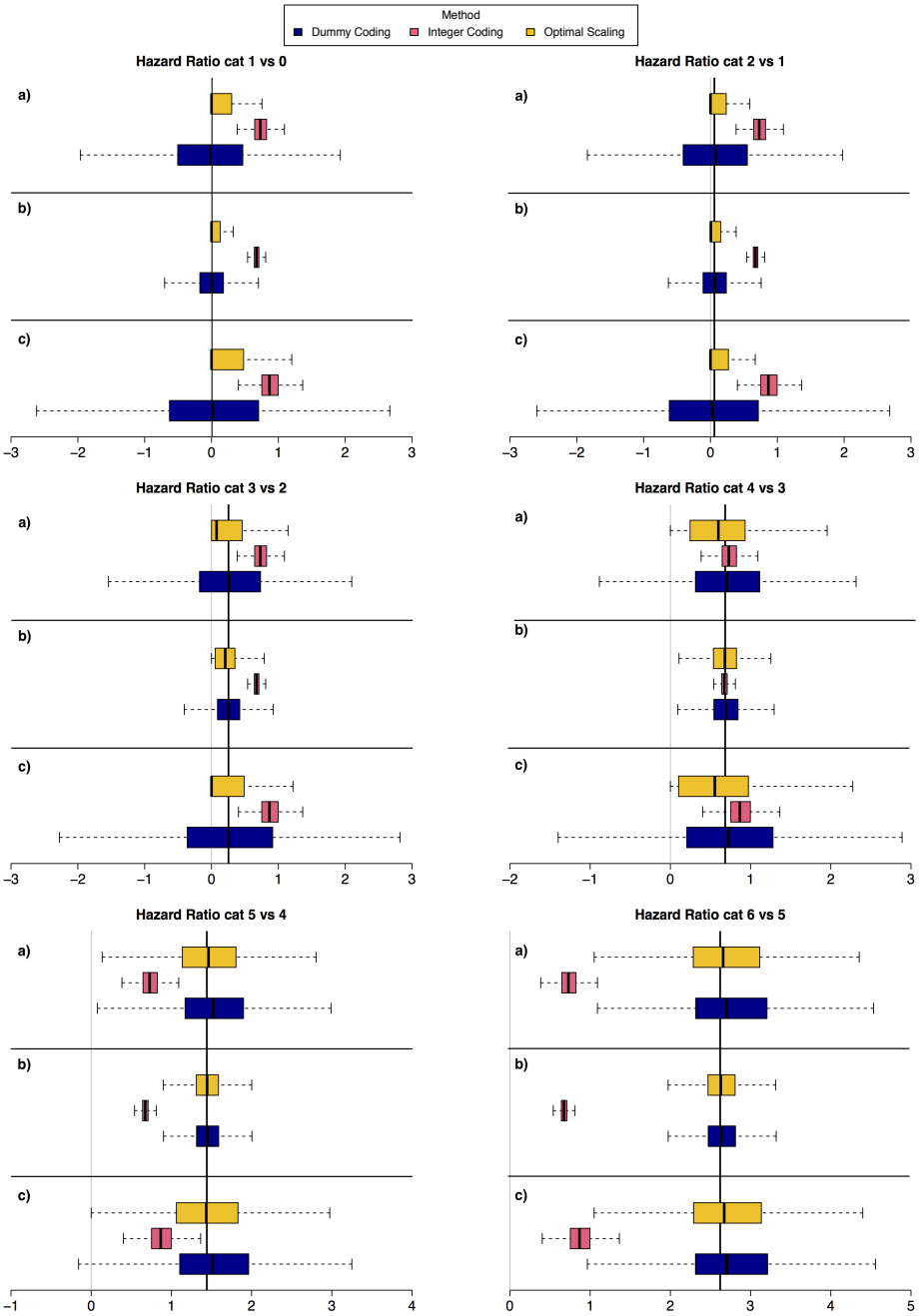


Figure 3.5: *Box plots of estimated log hazard ratios between consecutive categories based on the three methods (see Table 3.7) for each of the three scenarios; a) Scenario 1, b) Scenario 2, and c) Scenario 3. Simulation parameters for each scenario were chosen as in Table 3.8. The black vertical lines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

3.5 Discussion

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

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

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

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

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

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

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

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

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

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

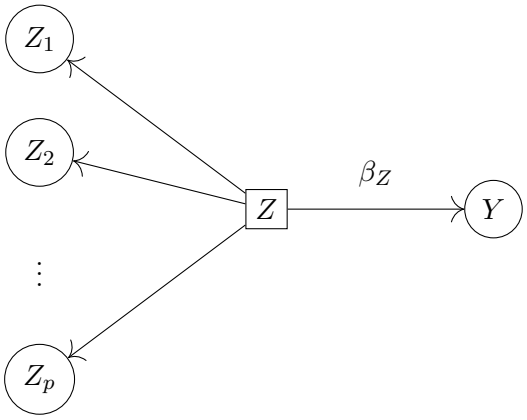


Figure 3.6: *Representation of dimension reduction into one dimension for the regression model.*

3.6 Supplementary material

3.6.1 Optimal scaling in ordinary linear regression

Recall the loss function

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

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

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

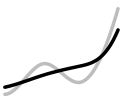
with $D_k = \mathbf{G}_k^T \mathbf{G}_k$, the diagonal matrix that gives the number of objects in each category. Similarly, the ordinary least squares solution can be used to minimize loss function (3.36) over β_k while keeping $\boldsymbol{v}_k$ fixed. i.e.

$$\begin{aligned}
\widehat{\beta}_k &= (\widehat{\varphi}_k(\boldsymbol{x}_k)^T \widehat{\varphi}_k(\boldsymbol{x}_k))^{-1} \widehat{\varphi}_k(\boldsymbol{x}_k)^T \boldsymbol{u}_k \\
&= (\widehat{\varphi}_k(x_{k1})^2 + \ldots + \widehat{\varphi}_k(x_{kp})^2)^{-1} \widehat{\varphi}_k(\boldsymbol{x}_k)^T \boldsymbol{u}_k \\
&= (||\widehat{\varphi}_k(\boldsymbol{x}_k)||^2)^{-1} \widehat{\varphi}_k(\boldsymbol{x}_k)^T \boldsymbol{u}_k \\
&= 1^{-1} \widehat{\varphi}_k(\boldsymbol{x}_k)^T \boldsymbol{u}_k \\
&= \boldsymbol{u}_k^T \widehat{\varphi}_k(\boldsymbol{x}_k).
\end{aligned} \tag{3.37}$$

3.6.2 From maximum likelihood to least squares

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

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



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

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

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

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

$$\begin{aligned}
& \tfrac{1}{2}(\zeta(\widetilde{\boldsymbol{\eta}}) - \mathbf{Z}\boldsymbol{\beta})^T\ell''(\widetilde{\boldsymbol{\eta}})(\zeta(\widetilde{\boldsymbol{\eta}}) - \mathbf{Z}\boldsymbol{\beta}) \\
= \;& \tfrac{1}{2}(\widetilde{\boldsymbol{\eta}} - \ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}) - \mathbf{Z}\boldsymbol{\beta})^T\ell''(\widetilde{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\eta}} - \ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}) - \mathbf{Z}\boldsymbol{\beta}) \\
= \;& \tfrac{1}{2}(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta} - \ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))^T\ell''(\widetilde{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta} - \ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}})) \\
= \;& \tfrac{1}{2}\big[(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta})^T\ell''(\widetilde{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta}) - (\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta})^T\ell''(\widetilde{\boldsymbol{\eta}})(\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}})) \\
& -(\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))^T\ell''(\widetilde{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta}) + (\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))^T\ell''(\widetilde{\boldsymbol{\eta}})(\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))\big] \\
= \;& \tfrac{1}{2}\big[(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta})^T\ell''(\widetilde{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta}) - (\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta})^T\ell'(\widetilde{\boldsymbol{\eta}}) \\
& -\ell'(\widetilde{\boldsymbol{\eta}})^T(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta}) + (\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))^T\ell'(\widetilde{\boldsymbol{\eta}})\big] \\
= \;& \tfrac{1}{2}\big[(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta})^T\ell''(\widetilde{\boldsymbol{\eta}})(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta}) - 2(\widetilde{\boldsymbol{\eta}} - \mathbf{Z}\boldsymbol{\beta})^T\ell'(\widetilde{\boldsymbol{\eta}}) + (\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))^T\ell'(\widetilde{\boldsymbol{\eta}})\big] \\
= \;& \tfrac{1}{2}\big[(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}})^T\ell''(\widetilde{\boldsymbol{\eta}})(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}}) + 2(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}})^T\ell'(\widetilde{\boldsymbol{\eta}}) + (\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))^T\ell'(\widetilde{\boldsymbol{\eta}})\big] \\
= \;& \tfrac{1}{2}(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}})^T\ell''(\widetilde{\boldsymbol{\eta}})(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}}) + (\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}})^T\ell'(\widetilde{\boldsymbol{\eta}}) + (\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))^T\ell'(\widetilde{\boldsymbol{\eta}}) \\
= \;& (\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}})^T\ell'(\widetilde{\boldsymbol{\eta}}) + \tfrac{1}{2}(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}})^T\ell''(\widetilde{\boldsymbol{\eta}})(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}}) + (\ell''(\widetilde{\boldsymbol{\eta}})^{-1}\ell'(\widetilde{\boldsymbol{\eta}}))^T\ell'(\widetilde{\boldsymbol{\eta}}) \\
= \;& (\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}})^T\ell'(\widetilde{\boldsymbol{\eta}}) + \tfrac{1}{2}(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}})^T\ell''(\widetilde{\boldsymbol{\eta}})(\mathbf{Z}\boldsymbol{\beta} - \widetilde{\boldsymbol{\eta}}) + C(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\eta}}).
\end{aligned}$$

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

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

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

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

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

3.6.3 Derivatives log likelihood Cox model

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

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

The log likelihood is

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

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

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

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

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