Advances in Survival Analysis and Optimal Scaling Methods

Willems, S.J.W.

Introduction

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

Survival analysis

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

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

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

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

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

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

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

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

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

which is a constant.

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

Optimal scaling transformations

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

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

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

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

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

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

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

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

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

The nominal and ordinal scaling levels are usually applied to categorical data. The former only preserves the grouping property of a predictor, and the latter preserves both the grouping and ordering properties. For continuous data or categorical variables with many category levels, a smooth spline transformation is more suitable. Spline functions allow for nonlinear transformations which can either preserve the ordering property of the data (monotone spline) or not (nonmonotone spline). If a linear relation is expected, linearity constrains can be applied by choosing the numeric scaling level; this will give the same result as in linear regression.

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

Statistics communication

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

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

Extensive research has been done on the interpretation of English verbal probability phrases. This research showed that there is huge variation in the interpretation of verbal probability phrases and that interpretation is often asymmetric. For example, the mean perceived percentages of mirrored expressions as *likely* and *unlikely* do not sum to 100% (Lichtenstein and Newman, 1967; Reagan et al., 1989; Stheeman et al., 1993). Additionally, interpretation is influenced by the base rate expectation of the statement in which the phrase is placed. For example, the numerical interpretation of *likely* in the statement *"It is likely that it will rain in Manchester, England, next June"* is usually higher than in *"It is* *likely that it will rain in Barcelona, Spain, next June"* (Wallsten et al., 1986).

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

Outline of this thesis

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

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

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

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

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