Advances in Survival Analysis and Optimal Scaling Methods

Willems, S.J.W.

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

4

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

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

4.1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

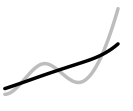
The type of transformation (also called *scaling level* in the categorical data analysis context) is chosen for each individual variable and may thus differ among predictors. The combination of coefficients and transformations calculated by the algorithm optimally describe the relation between the response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

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

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

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

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



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

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

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

4.2 Optimal scaling in linear regression

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

4.2.1 Model and notation

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

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

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

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

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

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

4.2.2 Model estimation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

OS-regression algorithm:
Initialization: Create $\mathbf{G}_1, \ldots, \mathbf{G}_p$ based on the data, and initialize the model parameters $\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p$ and $\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_p$.
Cycle: For $k = 1, \ldots, p$, do:
Step 1: Calculate $\mathbf{u}_k = \mathbf{y} - \sum_{l \neq k} \beta_l \mathbf{G}_l \mathbf{v}_l$.
Step 2: If the scaling level of variable k is nonnumeric, calculate the unrestricted estimates of the quantifications of k as

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

Apply appropriate scaling restrictions to $\widetilde{\mathbf{v}}_k^+$ and standardize the result.
Step 3: Update the estimate for model coefficient β_k as

$$\widetilde{\beta}_k^+ = \left\{ \widetilde{\mathbf{v}}_k^T \boldsymbol{D}_k \widetilde{\mathbf{v}}_k \right\}^{-1} \widetilde{\mathbf{v}}_k^T \mathbf{G}_k^T \mathbf{u}_k.$$

Convergence: Repeat the cycle until convergence criteria are met.

4.3 Optimal scaling in generalized linear models

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

4.3.1 GLM-OS model and notation

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

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

A GLM consists of three components, namely

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

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

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

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

4.3.2 Model estimation

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

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

Newton-Raphson method for GLMs

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

follows

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

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

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

$$\widetilde{\boldsymbol{\beta}}^{+} = \widetilde{\boldsymbol{\beta}} - \mathbf{H}_l^{-1}(\widetilde{\boldsymbol{\beta}})\,\boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\beta}}). \tag{4.7}$$

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

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

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

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

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

$$\boldsymbol{\nabla}_l(\mathbf{v}_1) = \ldots = \boldsymbol{\nabla}_l(\mathbf{v}_p) = \mathbf{0},$$

and

$$\boldsymbol{\nabla}_l(\beta_0) = \ldots = \boldsymbol{\nabla}_l(\beta_p) = 0.$$

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

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

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

where

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

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

$$\boldsymbol{\nabla}_l(\widetilde{\mathbf{v}}_k) = (\widetilde{\beta}_k \mathbf{G}_k)^T\ \boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\eta}})$$

and

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

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

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

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

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

$$\widetilde{\beta}_k^+ = \widetilde{\beta}_k - \mathbf{H}_l^{-1}(\widetilde{\beta}_k)\ \boldsymbol{\nabla}_l(\widetilde{\beta}_k), \tag{4.9}$$

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

Using the chain rule,

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

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

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

GLM-OS algorithm:
Initialization: Set $\mathbf{G}_0 = \mathbf{1}_n$ and $\mathbf{v}_0 = \{1\}$, create $\mathbf{G}_1, \ldots, \mathbf{G}_p$ based on the data, and initialize the model parameters $\widetilde{\beta}_0, \ldots, \widetilde{\beta}_p$ and $\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_p$.
Cycle: For $k = 0, \ldots, p$, do:
Step 1: Calculate the Hessian matrix $\mathbf{H}_l(\widetilde{\boldsymbol{\eta}})$ and the gradient vector $\boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\eta}})$.
Step 2: If the scaling level of variable k is nonnumeric, calculate the unrestricted estimates of the quantifications of k as

$$\widetilde{\mathbf{v}}_k^+ = \widetilde{\mathbf{v}}_k - \left\{(\widetilde{\beta}_k\mathbf{G}_k)^T\ \mathbf{H}_l(\widetilde{\boldsymbol{\eta}})\ \widetilde{\beta}_k\mathbf{G}_k\right\}^{-1}\ (\widetilde{\beta}_k\mathbf{G}_k)^T\ \boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\eta}}).$$

Apply appropriate scaling restrictions to $\widetilde{\mathbf{v}}_k^+$ and standardize the result.
Step 3: Update the Hessian matrix $\mathbf{H}_l(\widetilde{\boldsymbol{\eta}})$ and the gradient vector $\boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\eta}})$ using the current estimate $\widetilde{\mathbf{v}}_k$.
Step 4: Update the estimate for model coefficient β_k as

$$\widetilde{\beta}_k^+ = \widetilde{\beta}_k - \left\{(\mathbf{G}_k\widetilde{\mathbf{v}}_k)^T\ \mathbf{H}_l(\boldsymbol{\eta})\ \mathbf{G}_k\widetilde{\mathbf{v}}_k\right\}^{-1}\ (\mathbf{G}_k\widetilde{\mathbf{v}}_k)^T\ \boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\eta}}).$$

Convergence: Repeat the cycle until convergence criteria are met.

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

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

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

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

$$\widetilde{\boldsymbol{\beta}}^{+} = \widetilde{\boldsymbol{\beta}} - \mathbf{H}_l^{-1}(\widetilde{\boldsymbol{\beta}})\ \boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\beta}}).$$

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

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

and

$$\boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\beta}}) = \mathbf{X}^T\ \boldsymbol{\nabla}_l(\widetilde{\boldsymbol{\eta}}).$$

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

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

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

$$\text{argmin}_{\boldsymbol{\beta}}\ \left\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\right\|^2_{\mathbf{H}_l(\widetilde{\boldsymbol{\eta}})},$$

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

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

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

and

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

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

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

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

4.3.4 Example: logistic regression with optimal scaling transformations

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

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

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

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

with corresponding log-likelihood

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

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

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

$$\boldsymbol{\nabla}_{-l}(\widetilde{\boldsymbol{\eta}}) = \boldsymbol{\pi} - \mathbf{y}$$

and the Hessian is

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

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

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

$$\widetilde{\mathbf{v}}_k^{+} = \widetilde{\mathbf{v}}_k - \left[(\widetilde{\beta}_k\mathbf{G}_k)^T \operatorname{diag}\{\boldsymbol{\pi}(1-\boldsymbol{\pi})\}\ \widetilde{\beta}_k\mathbf{G}_k\right]^{-1}\ (\widetilde{\beta}_k\mathbf{G}_k)^T\,(\boldsymbol{\pi}-\mathbf{y}) \qquad (4.15)$$

and

$$\widetilde{\beta}_k^{+} = \widetilde{\beta}_k - \left[(\mathbf{G}_k\widetilde{\mathbf{v}}_k)^T \operatorname{diag}\{\boldsymbol{\pi}(1-\boldsymbol{\pi})\}\ \mathbf{G}_k\widetilde{\mathbf{v}}_k\right]^{-1}\ (\mathbf{G}_k\widetilde{\mathbf{v}}_k)^T\,(\boldsymbol{\pi}-\mathbf{y}), \qquad (4.16)$$

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

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

4.4 Application of GLM with optimal scaling: logistic regression

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

4.4.1 Transformation and visualization of categorical predictors

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

Data description

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

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

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

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

OS transformations with nominal scaling level

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

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

Category	Dummy coding	Optimal scaling
1		$\beta_k v_{k_1}$
2	β_{k_2}	$\beta_k v_{k_2}$
⋮	⋮	⋮
$C_k - 1$	$\beta_{k_{C_k-1}}$	$\beta_k v_{k_{C_k-1}}$
C_k	$\beta_{k_{C_k}}$	$\beta_k v_{k_{C_k}}$

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

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

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

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

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

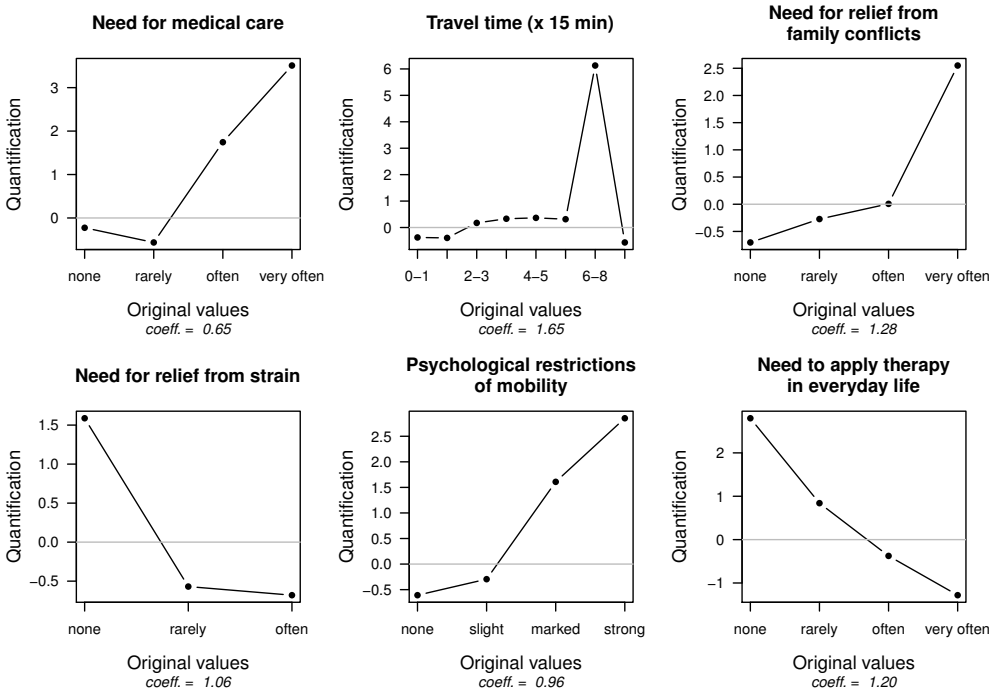


Figure 4.1: *DINSTAP data: Nominal quantifications estimated for each of the original categories of the six predictor variables. Regression coefficients are provided below the plots of the corresponding predictor. The estimated intercept is 0.56.*

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

Concluding, the visualizations of the quantifications help to interpret the

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

Comparison between OS transformations and the use of dummy variables

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

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

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

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

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

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

4.4.2 Monotone transformations to facilitate interpretation

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

Data description

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

Nonmonotone vs. monotone quantifications

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

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

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

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

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

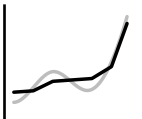
Hence, although differences with the nonmonotonic results are small, the monotonic quantifications of *Education Husband*, *Occupation Husband*, and *Number of previous children* are easier to interpret and correspond more to reality than the nonmonotonic ones.

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

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

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

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



Relation with ordinary logistic regression and GAMs

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

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

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

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

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

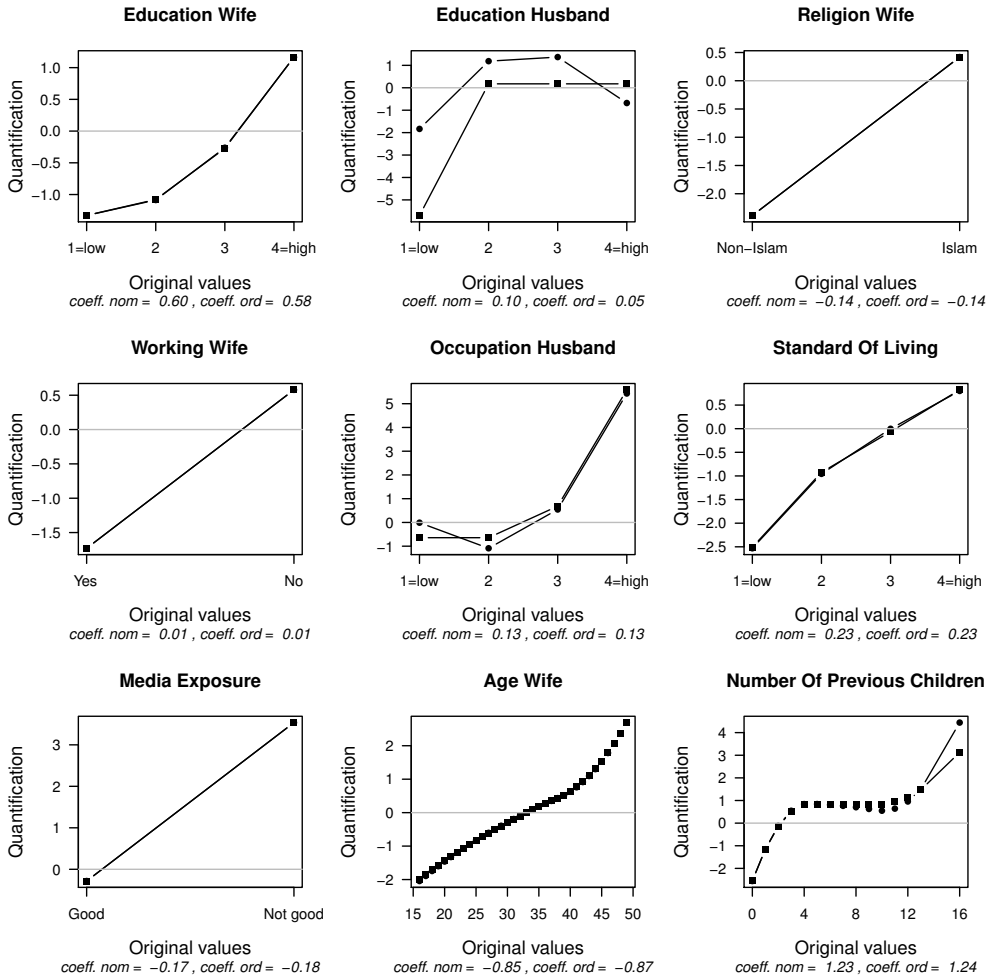


Figure 4.2: *Contraceptive method choice data: Nominal (circles) and ordinal (squares) quantifications estimated for each of the original categories of the nine predictor variables. Categorical variables are transformed using step functions and continues predictors are transformed using splines. Estimated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

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

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

4.4.3 Mixed scaling levels

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

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

Data description

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

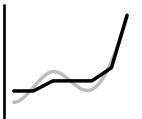
The predictor variables were measured on different scales. Variables *Node caps*, *Irradiation*, and *Breast* are categorical with two unordered categories. *Breast quadrant* and *Menopause* are categorical with more than two unordered categories. The *Degree of malignancy* is indicated by three category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

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

GLM-OS with scaling levels according to measurement level

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

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



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

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

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

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

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

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

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

Comparison with nonmonotone scaling level and linearity restrictions

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

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

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

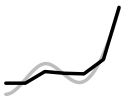
Table 4.5: *Breast cancer recurrence data: Apparent prediction error (APE) for the GLM-OS model with nonmonotone and monotone transformations, together with the 10-fold cross validation: Expected Prediction Error (EPE) along with its standard error (SE(EPE)) and the Misclassification Rate (MCR). Results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

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

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

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

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



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

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

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

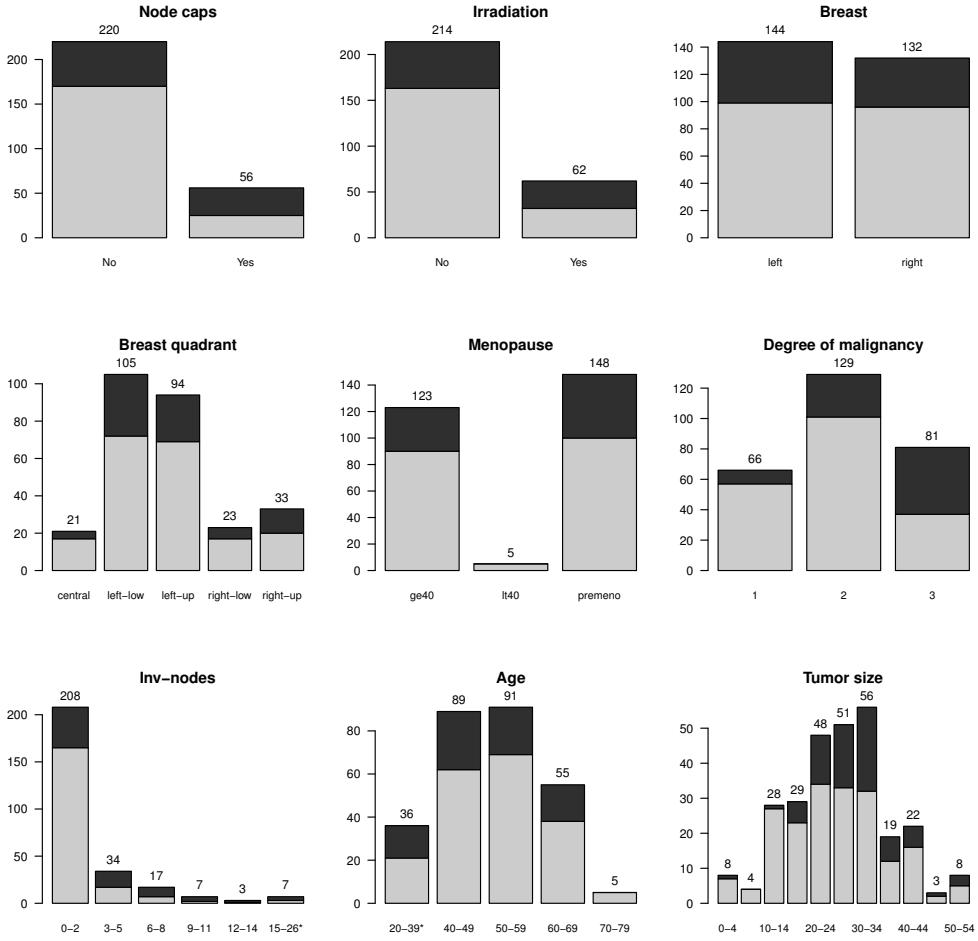


Figure 4.3: *Breast cancer recurrence data: Distribution of observation over the categories of the predictor variables, split by outcome value (dark grey = recurrence events, light grey = no recurrence events).*
**This category is a union of two categories that were merged because one of the original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

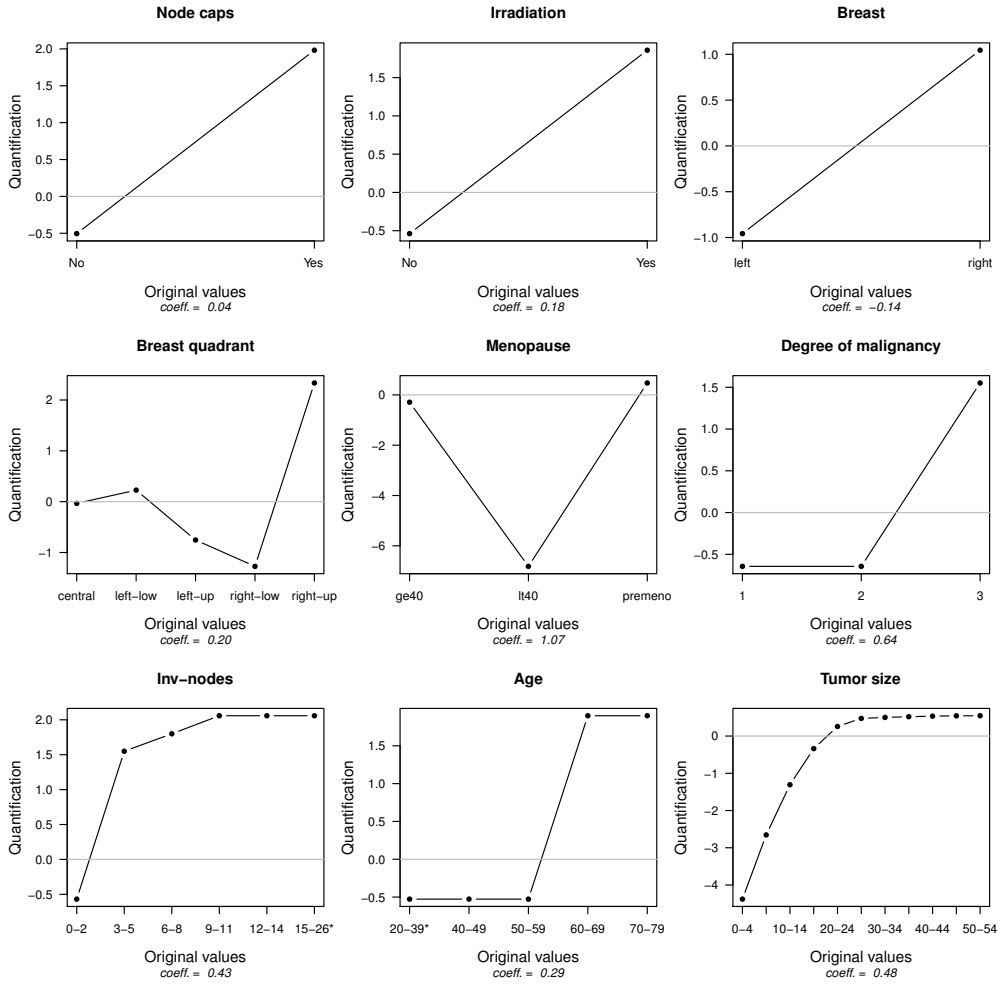


Figure 4.4: *Breast cancer recurrence data: Quantifications estimated for each of the original categories of the nine predictor variables. Unordered categorical variables are transformed using nonmonotone step functions and ordered categorical variables are transformed using monotone step functions. Variable* Tumor size *is transformed using a monotone spline. Regression coefficients are provided below the plots of the corresponding predictor. The estimated intercept is -1.24.*
**This category is a union of two categories that were merged because one of the original categories contained only one observation.*

4.5 Discussion

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

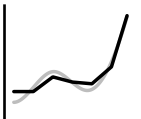
Transformations of the predictor variables have been integrated in GLMs before (Hastie and Tibshirani, 1990). However, the OS methodology has several benefits compared to other methods.

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

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

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

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



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

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

Acknowledgements

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

4.6 Supplementary material

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

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

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

Using this representation, the probability distribution for Y_i is

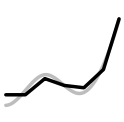
$$p(y_i) = P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}.$$

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

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

and the corresponding log-likelihood function is

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



To simplify computations the negative log-likelihood

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

is minimized. The gradient of l^{-} is the vector with elements

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

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

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

In matrix notation,

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

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