Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Peter Rusnák

Zobecněné lineární a aditivní modely v pojišťovnictví

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Martin Branda, Ph.D.
Studijní program: Matematika
Studijní obor: MFAPM

Praha 2013

Charles University in Praque
Faculty of Mathematics and Physics

MASTER THESIS

Peter Rusnák

Generalized and additive models in insurance

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Martin Branda, Ph.D.

Study programme: Mathematics

Specialization: MFAPM

Prague 2013

I would like to thank my supervisor RNDr. Martin Branda, Ph.D. for precise guidance and useful advice. I also wish to thank Mgr. Renáta Ševčíková for helpful consultations and support. My gratitude goes to ℝ-software creators and Generali Pojišťovna, a.s., which supported this thesis. Finally, I would like to thank my family and friends for support and companionship during my studies.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague 2.8.2013 .....

Název práce: Zobecněné lineární a aditivní modely v pojišťovnictví

Autor: Peter Rusnák

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Martin Branda, Ph.D.

Abstrakt: V předložené práci se věnujeme teorii zobecněných lineárních modelů a jejích aplikacím v oblasti pojišťovnictví. Představíme také některé metody běžně používané pro odhad regresních parametrů a testování hypotéz. Zaměříme se na možné rozšíření GLM zavedením nástroje pro parametrizaci prediktorů, která povede k nové třídě modelů, konkrétně odvodíme segmentované zobecněné lineární modely a zobecněné aditivní modely.V práci také uvádíme modely vhodné pro aktuárskou praxi. Nakonec v praktické části této práce ilustrujeme použití vhodného softwaru pro výpočet parametrů GAM a představujeme způsob, jak využít open source statistický program ℝ.

Klíčová slova: regrese, linearní model, odhad paramterů, spline, pojišťovniství

Title: Generalized and additive models in insurance

Author: Peter Runsák

Department: Department of Probability and Mathematical Statistics

Supervisor of master thesis: RNDr. Martin Branda, Ph.D.

Abstract: In this thesis we describe the theory of generalized linear models and demonstrate its applications in non-life insurance. We also introduce some methods commonly used to estimation of regression parameters and hypothesis testing . Furthermore, we discuss possible extensions of GLM by introducing tools for reparametrization of predictors which leads to new classes of models, concretely to segmented generalized linear models and generalized additive models. Consequently, we derive models appropriate for actuarial praxis using the real insurance data. In practical part of this thesis we illustrate the use of appropriate software for calculating the parameters of GAM and find way how to use open source statistical program ℝ.

Keyword: regression, linear model, estimation of parameters, spline, insurance

Contents

Introduction

Regression analysis plays important role in statistics being one of its most powerful and commonly used tools for analysing relationships among variables in datasets. It includes many different techniques, from which only small portion is presented in the thesis, where the main focus always lies on the relationship between a response and one or more predictors.

The goal of this thesis is to describe and examine some largely used regression techniques from the theoretical point of view, but first and foremost demostrate this knowledge on practical example from non-life insurance.

In the first chapter we begin by introducing the components of the basic technique known as linear regression for the standard linear model (SLM). Although, we focus on more complicated techniques, it would be easier for us to understand them when we firstly describe the components of the one from which they all originate. Also we demonstrate that the linear regression is not always the proper choice due to big restrictions on the form of examined relationships, which justify latter introduced techniques.

Consequently, in the second chapter we introduce the Generalized Linear Model (GLM), which represent an extension of the SLM by enabling the regression analysis to deal with a wider class of relationships. Namely, we allow the response to have other than normal distribution and enable a degree of nonlinearity in the model structure. We discuss the key elements of GLM as well as available response distributions, link functions, maximum likelihood estimates and the fact that model fitting has to be done iteratively. Also, we show that the cost of such generalization is that distributional results are now approximate and justified by large sample limiting results, rather than being exact.

In the third chapter we explain new task to which we often refer in practical part of this thesis with real life data. Namely, that we cannot use one uniform function as proper transformation of predictor since the functional relationship between the mean of the response and the predictors changes at certain points of their domains. Hence, in this chapter we propose as proper technique Segmented Generalized Linear model, which extend the possibilities of GLM by enabling these relationships to be piecewise linear. Furthermore we introduce possible algorithms for localization of the points in which the relationship changes, called breakpoints.

We return to the problem of proper parametrization in the fourth chapter. However, we assume that the piecewise linear parameterization is not flexible enough to produce reliable models. Hence, we introduce Generalized Additive Model (GAM), which

extend possibilities of GLM by using penalized regression splines as reparametrization tool. This new flexibility and convenience yields to two new problems; it will be necessary not only to pay attention how represent these splines in some predefined way but also to properly reconsider and choose how smooth in the result these splines should be.

Finally, in the last fifth chapter we focus on practical demonstrations of mentioned models in terms of non-life insurance. We present business case for which satisfying solution we will have to use all introduced techniques. Because we threat this problem also as real life situation, we pay a big attention besides presenting modeling techniques also to business and interpretation side.

1. Preliminaries

We assume that the reader is familiar with the basics in regression. However, let us remind that the purposse of regression is to model and study the relationships between a given set of variables (predictors), for which one knows their true values (or is able to predict them), and a variable (response), in which estimated values is one interested.

We perform regression on dataset of observations which include the values of predictors as well as the value of response for particular observation.

Such dataset can be easily interpreted by:

- The matrix $\mathbb{X}$ representing the predictors

$$\mathbb{X} = \begin{pmatrix} X_{1,0} & \dots & X_{1,k} \\ \vdots & \ddots & \vdots \\ X_{n,0} & \cdots & X_{n,k} \end{pmatrix}$$

  for which we assume:

  - the predictor $X_{i,j}$ is a random variable uncontaminated with measurement errors. Consequently, it can be treated as some fixed value $x_{i,j}$ repressenting the i^{th} observation of j^{th} predictor,
  - the first predictor serves to model intercept and so $\forall\, i = 1, \dots, n$ it holds $x_{i,0} = 1$,
  - rank of the matrix $\mathbb{X}$ equals $k + 1$ (= the number of columns).

Note: These three assumptions are not necessery but dropping them leads to significantly more difficult models. Therefore, from now on we will use the following form of the predictor matrix:

$$\mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix}.$$

- The vectors representing the response:

$$\mathbb{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

  for which we assume:

  - meassurment y_i is the observation of corresponding random variable Y_i,
  - the distribution of the random variable Y_i depends on the matrix row $\mathbb{X}_i$,
  - vectors $Y_1, \dots, Y_n$ are mutually independent random variables.

In next two sections we introduce the basic regression model, Standard linear model, and we look closer at the estimation of regression coefficient for this model.

Because all definitions used in this chapter are contained in [7], the source is men-

tioned only if it is necessary or recommended.

1.1. Standard Linear Model (SLM)

In case of the Stanard Linear Model (SLM) we assume that the relationship between predictors and response is linear. And so, we are using linear regression to determining the corresponding relationship which can be expressed by following equation:

$$\mu_i = \mathbb{X}_{i.}\boldsymbol{\beta}, \quad \text{for } i \in \{1, \ ..., n\},$$

where

$$Y_i = \mu_i + \epsilon_i, \ \text{for } i \in \{1, \ ..., n\},$$
$$E\, Y_i = \mu_i, \ \text{for } i \in \{1, \ ..., n\}$$

and for errors (residuals) ϵ_i hold assumptions:

(E1) $\epsilon_1, \ ..., \epsilon_n$ are mutually independendent random variables,

(E2) $E(\epsilon_i) = 0$,

(E3) Constant variance: $var(\epsilon_i) = \sigma^2 > 0$.

Note: The variable $\boldsymbol{\beta}$ represents the vector of regression coefficients which values are unknown and our goal is to estimate them. Subsequently, based on this estimations we can determine the estimates of response itself.

Matrix form of model is

$$\boldsymbol{\mu} = \mathbb{X}.\boldsymbol{\beta},$$

where

$$\mathbb{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon},$$
$$E\,\mathbb{Y} = \boldsymbol{\mu},$$
$$\boldsymbol{\epsilon} = (\epsilon_1, \ ..., \epsilon_n)^T, \ \boldsymbol{\mu} = (\mu_1, \ ..., \mu_n)^T, \ \boldsymbol{\beta} = (\beta_0, \ ..., \beta_k)^T.$$

Since predictors are treated as fixed values, using their non-linear transformations does not cause any troubles. We list some common versions of SLM, where the predictors values are transformed:

1. *Model of the relationship between Y_i and $log(x_i)$*, where $x_i>0$

$$Y_i = \beta_0 + \beta_1 \cdot log\,(x_i) + \epsilon_i, \ \forall\ i \in \{1, \ ..., n\},$$

is

$$\mu_i = \beta_0 + \beta_1 \cdot log\,(x_i)\,, \ \forall\ i \in \{1, \ ..., n\}.$$

Therefore $k = 1$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$.

2. *Cubic model of the relationship between Y_i and x_i*

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \beta_3 \cdot x_i^3 + \epsilon_i, \ \forall\, i \in \{1, \ ..., n\},$$

is

$$\mu_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \beta_3 \cdot x_i^3, \ \forall\, i \in \{1, \ ..., n\}.$$

Therefore $k = 3$ and $\boldsymbol{\beta} = (\beta_0, \ ..., \beta_3)^T$.

1.2. The Least Square Estimation method

Let us introduce this basic method for estimation of regression coefficients $\boldsymbol{\beta}$ based on n-observations in linear regression.

Firstly we define some terms commonly used in statistics. Suppose, we have the observations $y_1,...,y_n$ of random variables $Y_1,...,Y_n$ with distribution functions which depend on some parameter $\boldsymbol{\beta}$ (vector). And we want to estimate this parameter $\boldsymbol{\beta}$.

<u>*Definition*</u>: A *statistic T* is any measurable function of random variables $Y_1,...,Y_n$; it is also a random variable. A *(point) estimator* is any statistic $T(Y_1, \ ..., \ Y_n)$. A *(point) estimate* is a realized value of an point estimator that is obtained when a sample of observations is already taken.

To estimate the regression coefficients $\boldsymbol{\beta}$ using Least Square Estimation (LSE) we use the following notation:

- vector $\hat{\boldsymbol{\beta}}_{\mathbb{y}}$- estimate of $\boldsymbol{\beta}$ for realization $\mathbb{y}$ of random sample $\mathbb{Y}$; computed using LSE:

$$\hat{\boldsymbol{\beta}}_{\mathbb{y}} = arg\ min_{\beta \in \mathbb{R}^{k+1}} (\mathbb{y} - \mathbb{X}\boldsymbol{\beta})^T (\mathbb{y} - \mathbb{X}\boldsymbol{\beta}) = \left(\mathbb{X}^T \mathbb{X}\right)^{-1} \mathbb{X}^T \mathbb{y}.$$

- vector $\hat{\boldsymbol{\mu}}_{\mathbb{y}}$ - estimate of $\boldsymbol{\mu}$ for realization $\mathbb{y}$ of random sample $\mathbb{Y}$ representing fitted values of response in LSE; computed as:

$$\hat{\boldsymbol{\mu}}_{\mathbb{y}} = \hat{\boldsymbol{\beta}}_{\mathbb{y}} \mathbb{X}.$$

- vector $\hat{\boldsymbol{\epsilon}}_{\mathbb{y}}$ - estimate of residuals $\boldsymbol{\epsilon}$ for realization $\mathbb{y}$ of random sample $\mathbb{Y}$ representing the difference between the fitted value of the response predicted by the model and the true data value of the response; computed as:

$$\hat{\boldsymbol{\epsilon}}_{\mathbb{y}} = \mathbb{y} - \hat{\boldsymbol{\mu}}_{\mathbb{y}}.$$

- descriptive statistic *RSS* (Residual Sum of Squares) - measures the discrepancy between the data and the estimations. Small value of *RSS* indicates a good fit of the model to the data:

$$RSS = \hat{\boldsymbol{\epsilon}}_{\mathbb{y}}^{\ T} \hat{\boldsymbol{\epsilon}}_{\mathbb{y}} = \left(\mathbb{y} - \hat{\boldsymbol{\mu}}_{\mathbb{y}}\right)^T \left(\mathbb{y} - \hat{\boldsymbol{\mu}}_{\mathbb{y}}\right).$$

The values $\hat{\boldsymbol{\beta}}_{\mathbb{y}}$ have to make the model best fitting to the data in the sense of

minimizing *RSS*.

- descriptive statistic S^2 (Residual variance) - represents estimate of σ^2 for realization $\mathbb{y}$ of random sample $\mathbb{Y}$:

$$S^2 = \frac{RSS}{n - \mathrm{Rank}(\mathbb{X})} = \frac{RSS}{n - (k+1)}.$$

To evaluate the reliability of such estimates we have to consider some properties of corresponding estimators. So we look at the estimators of $\boldsymbol{\beta}$ instead of estimates**:**

- $\hat{\boldsymbol{\beta}}$ - estimator of $\boldsymbol{\beta}$; expressed in the following way:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta \in \mathbb{R}^{k+1}} (\mathbb{Y} - \mathbb{X}\boldsymbol{\beta})^T (\mathbb{Y} - \mathbb{X}\boldsymbol{\beta}) = \left(\mathbb{X}^T \mathbb{X}\right)^{-1} \mathbb{X}^T \mathbb{Y}.$$

<u>*Note*</u>: By definition of statistic, we have $T(Y_1, \ldots, Y_n) = LSE(Y_1, \ldots, Y_n) = \hat{\beta}$.

In [7] is shown, that the estimator $\hat{\boldsymbol{\beta}}$ has following attributes:

i. $E\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$,

ii. $\mathrm{var}\,\hat{\boldsymbol{\beta}} = \sigma^2\left(\mathbb{X}^T \mathbb{X}\right)^{-1}$,

iii. $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, consistent estimate.

By Gauss–Markov theorem ([7]) can be proven, that such estimator $\hat{\boldsymbol{\beta}}$ under introduced valid conditions is the Best Linear Unbiased Estimator (BLUE) of the regression coefficients $\boldsymbol{\beta}$ and so it holds:

iv. $E\left(\sum_{j=1}^{k+1} \left(\hat{\beta}_j - \beta_j\right)^2\right)$ is minimal.

- $\hat{\boldsymbol{\mu}}$ - point estimator of $\boldsymbol{\mu}$:

$$\hat{\boldsymbol{\mu}} = \mathbb{X}.\hat{\boldsymbol{\beta}},$$

by applying properties of $\hat{\boldsymbol{\beta}}$ we get:

i. $E\,\hat{\boldsymbol{\mu}} = \mathbb{X}.\boldsymbol{\beta} = \boldsymbol{\mu}$,

ii. $\mathrm{var}\,\hat{\boldsymbol{\mu}} = \sigma^2\, \mathbb{X}.\left(\mathbb{X}^T.\mathbb{X}\right)^{-1}.\mathbb{X}^T$,

iii. $\hat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$, consistent estimate.

- $\hat{\boldsymbol{\epsilon}}$ - estimator of $\boldsymbol{\epsilon}$:

$$\hat{\boldsymbol{\epsilon}} = \mathbb{Y} - \hat{\boldsymbol{\mu}},$$

by applying properties of $\hat{\boldsymbol{\beta}}$ we get:

i. $E\,\hat{\boldsymbol{\epsilon}} = 0$,

ii. $\mathrm{var}\,\hat{\boldsymbol{\epsilon}} = \sigma^2.\left(I_n - \mathbb{X}.\left(\mathbb{X}^T.\mathbb{X}\right)^{-1}.\mathbb{X}^T\right)$; I_n is the $n \times n$ identity matrix.

- $\hat{\sigma}^2$ - point estimator of σ^2:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}^T.\hat{\boldsymbol{\epsilon}}}{n - \text{Rank}(\mathbb{X})}.$$

Based on this properties of LSE $\hat{\boldsymbol{\beta}}$ and by adding assumption for distribution of the response we are able to evaluate the reliability of estimates by testing hypotheses related to the model. Usually, in SLM we consider the distribution to be normal, $\mathbb{Y} \sim N_n(\mu, \sigma^2.I_n)$. In this case it can be proven e.g. :

i. $\hat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\right)$,

ii. $\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-(k+1)}$,

iii. $\frac{\left(\hat{\beta}_i - \beta_i\right)}{\sqrt{\left(\sigma^2\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\right)_{i,i}}} \sim t_{n-(k+1)}$.

Different distributions can be replaced by asymptotic normal distribution for big number of observations:

iv. $\sqrt{n}.\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N_{k+1}\left(0, \sigma^2\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\right)$.

Note: Formulations of all used claims together with their proofs can be found in [7].

Standard linear model predicts the expected value of the response variable as a linear combination of predictors. This implies that a constant change in the predictors leads to a constant change in the response variable which is appropriate when a response variable can vary essentially randomly in either direction by a relatively small amount as it is in normal distribution.

However, the assumption of linear relationship between the response and the predictors becomes very problematic for distributions which put constraints on allowable values of response. For example, if the response has binomial distribution, we expect fitted values to be in interval [0,1] but by using SLM we could encounter situations when the predicted values are for some values of predictors outside this interval. Hence, the assumptions in SLM reduce the practical usage of this model to only a few cases in which the response actually behaves in the requested way.

Possible way how to deal with this problem is to allow the response variable to depend on predictors through a non-linear function g :

$$g(Y_i) = \mathbb{X}_i.\boldsymbol{\beta} + \epsilon_i, \ \forall\, i \in \{1, \ldots n\}.$$

Such relationship could lead to the two possible models:

$$g(E(Y_i)) = \mathbb{X}_i.\boldsymbol{\beta}\,,\ \forall\ i \in \{1,\ \ldots n\}$$

and

$$E\,(g\,(\mathrm{Y_i})) = \mathbb{X}_\mathrm{i}.\boldsymbol{\beta},\ \forall i \in \{1,\ \ldots\ \mathrm{n}\}.$$

While the first one represents generalized linear model where the mean is transformed by the link function (Chapter 2.), the second one can be rewritten as SLM by transfroming the response and declaring new variable Z_i:

$$Z_i = g(Y_i)\,,\ \forall\ i \in \{1,\ \ldots n\}.$$

These two models can lead (recall Jensen inequality) to quite different results.

Note: For serie of observations it is not the same if we take log of average of these observations or average of log of observations.

And although, because of easier interpretation it may appear that the mean of the log-transformed response is prefferable, from a practical point of view is the log-transformed mean of response typically much more useful. The reason for this conclusion comes from recognition, that allowing the response variable Y_i to depend on predictors through a non-linear function g causes "much bigger problems" as we demonstrate in following example.

Example: Let us assume that $\epsilon_i \sim N\left(0,\ \sigma^2\right)$ and the non-linear function g is defined as log function. Then for $\forall\ Y_i > 0$ we have:

$$g(Y_i) = log(Y_i),\ \forall\ i \in \{1,\ \ldots n\}.$$

After transforming the response itself we get:

$$log(Y_i) = \mathbb{X}_i.\boldsymbol{\beta} + \epsilon_i,\ \forall\ i \in \{1,\ \ldots n\},$$

Subsequently, we can express LSE $\hat{\boldsymbol{\beta}}$ for model:

$$E(log(Y_i)) = \mathbb{X}_i.\boldsymbol{\beta},\ \forall\ i \in \{1,\ \ldots n\}.$$

and get:

$$E\left(\mathbb{X}_i.\hat{\boldsymbol{\beta}}\right) = \mathbb{X}_i.\boldsymbol{\beta}\ = \mu_i\ = E\,\hat{\boldsymbol{\mu}}_{\boldsymbol{i}},\ \forall\ i \in \{1,\ \ldots n\}.$$

By standard transformations of log-relationship between response and predictors we get:

$$Y_i = exp(E\,\hat{\mu}_{\boldsymbol{i}}).e^{\epsilon_i},\ \text{where}\ \epsilon_i \sim N\left(0,\ \sigma^2\right),$$

$$E(e^{\epsilon_i}) = \ e^{\frac{\sigma^2}{2}},\ \forall\ i \in \{1,\ \ldots n\}.$$

And so:

$$E\ Y_i = exp(E\,\hat{\mu}_{\boldsymbol{i}})\,e^{\frac{\sigma^2}{2}}\ ,\ \forall\ i \in \{1,\ \ldots n\}.$$

Looking at the above formula, we can see that the mean of response increases as variance of response increases.

Even more from:

$$\text{var}(e^{\epsilon_i}) = e^{\sigma^2} \cdot \left(e^{\sigma^2} - 1\right), \ \forall\, i \in \{1, \ \ldots n\}.$$

and using the assumption that the errors are uncorrelated with the predictor variables we get:

$$\text{var}(Y_i) = \exp(\mathbb{X}_i.\boldsymbol{\beta})^2.\, e^{\sigma^2} \cdot \left(e^{\sigma^2} - 1\right), \ \forall\, i \in \{1, \ \ldots n\}.$$

From the last equation we can see that the assumption of equal variances (E3) was broken.

These are the primary causes of the fact, that regression coefficients can not be estimated by presented LSE. The fitting of regression coefficients has to be done iteratively by some algorithm for non-linear least squares.

2. Generalized Linear Model (GLM)

By GLM can be fitted certain forms of non-linear models. The idea was formulated by John Nelder and Robert Wedderburn ([4]) as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression. They proposed an iteratively reweighted least squares method for maximum likelihood estimation of the model parameters. That allows us to consider models with other than linear dependece between predictors and mean of response and other type of response distribution than normal.

Three Keystones of GLM

A **random component** specifying the conditional distribution of the response variable, Y_i, depending on the values of the predictors in the model. In this work we are using Nelder and Wedderburn's original formulation that the distribution of Y_i is a member of an family of exponential distributions*;* so all corresponding definitions can be found in [4].

Note: There are methods which were developed to extend GLM to have the distribution of response variable from multivariate exponential family (such as the multinomial distribution) or certain nonexponential families (such as the two-parameter negative-binomial distribution) or there are methods for which the distribution of response variable is not specified completely, but this is not part of this thesis.

A **linear predictor** η_i is defined as linear function of predictors:

$$\eta_i = \mathbb{X}_i \boldsymbol{\beta}, \forall\, i \in \{1, \ldots n\}.$$

The structure of linear predictor reminds the structure of the standard linear model. And as in SLM we can use transformations of predictors to create extended versions of GLM.

A **link function** g, strictly monotonic and twice differentiable; transforms the mean of the response variable to the linear predictor:

$$g(E(Y_i)) = \mathbb{X}_i \boldsymbol{\beta}, \ \forall\, i \in \{1, \ldots n\}.$$

$$g(\mu_i) = \eta_i, \ \forall\, i \in \{1, \ldots n\}.$$

Note: Because the link function is strictly monotonic it is also invertible and we can write:

$$\mu_i = g^{-1}(\eta_i), \ \forall\, i \in \{1, \ldots n\}.$$

The inverse link g^{-1} is called the *mean function*.

To sum it up, suppose that an appropriate formula describing the relationship between $\mathbb{X}_i$ and Y_i is:

$$g(Y_i) = \mathbb{X}_i \boldsymbol{\beta} + \epsilon_i, \ \forall\, i \in \{1, \ \ldots n\},$$

then the basic structure of GLM is:

$$g(\mu_i) = \eta_i, \ \forall\, i \in \{1, \ \ldots n\}.$$

2.1. Family of exponential distributions

As was mentioned before, we make assumptions that Y_i, $i \in \{1, \ldots \text{n}\}$ are mutually independent random variables and distributions of Y_i, $i \in \{1, \ldots \text{n}\}$ belong to the family of exponential distributions. We also assume that the type of distribution of response is known. In this section we take a closer look at distributions from the family of exponential distribution.

Defintition ([7]): A distribution of random variable Y belongs to the *family of exponential distributions* (exponential family), if it's probability density function can be written as function:

$$f(y, \theta, \phi) = exp\left\{\frac{y\,\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

where a, b, c are known functions, ϕ is so called *scale parameter* and θ is so called *canonical parameter* of the distribution.

In this thesis we work only with distributions where the function a is defined as

$$a\,(\phi) \ = \ \phi\,/\,\omega,$$

where ω is known constant.

This restriction of definition of function a suffices for all practical examples of GLM presented in this thesis. The known constants ω represents weight of particular observation and we use it e.g. to compensate different time exposures of observations.

The exponential family of distribution includes many distributions that are useful for practical modelling, especially for usage in insurance are the Poisson, Gamma and Binomial very suited and common. Determination of functions a (only in restricted form), b, c and parameters θ, ϕ for these three distributions is subsequently demonstrated.

Note: Normal distribution also belongs to the exponential family of distribution.

- **Binomial distribution**

  - the values of Y belong to $\{1, \ldots, n\}$, where n represents number of observations.

  Density function for expected mean value $E(Y) = \mu$

$$f_\mu(y) = \binom{n}{y}\left(\frac{\mu}{n}\right)^y\left(1-\frac{\mu}{n}\right)^{n-y}$$

  can be rewritten as:

$$f_\mu(y) = \exp\left\{y\,log\left(\frac{\mu}{n-\mu}\right) - n\,log\left(1+\frac{\mu}{n-\mu}\right) + log\binom{n}{y}\right\}.$$

  The form of density function for exponential family of distributions can be achieved by following substitutions:

  $\phi = 1,$

  $a(\phi) = 1,$

  $\theta = log\left(\frac{\mu}{n-\mu}\right),$

  $b(\theta) = n \cdot log\left(1+e^\theta\right),$

  $c(y, \phi) = log\binom{n}{y}.$

  <u>*Note*</u>: Further we use special case of binomial distribution and alternative distribution for modelling of probability of specific occurance.

- **Poisson distribution**

  - the values of Y belong to $\mathbb{N}_0$.

  Density function for expected mean value $E(Y) = \mu$

$$f_\mu(y) = \frac{\mu^y\, exp(-\mu)}{y!}$$

  can be rewritten as:

$$f_\mu(y) = \exp\{y\,log(\mu) - \mu - log(y!)\}.$$

  The form of density function for expenential family of distributions can be achieved by following substitutions:

  $\phi = 1,$

  $a(\phi) = 1,$

  $\theta = log(\mu),$

  $b(\theta) = e^\theta,$

  $c(y, \phi) = -log(y!).$

  <u>*Note*</u>: Further we use Poisson distribution for modelling of number of specific occurances.

- **Gamma distribution**

  -the values of Y belong to the interval $(0, \infty)$.

  Density function for expected mean value $E(Y) = p / a$

$$f_{(p,a)}(y) = \frac{a^p \, y^{p-1} \, exp(-a \, y)}{\Gamma(p)}$$

  can be rewritten as:

$$f_{(p,a)}(y) = \exp\left\{\frac{y\,(-p/a) + log(-p/a)}{1/p} + p\log(p) - log\,\Gamma(p) + (p-1)\,log\,(y)\right\}.$$

  Form of density function from expenential family of distributions can be achieved by following substitutions:

$$\phi = p,$$

$$a(\phi) = 1/p,$$

$$\theta = -p/a,$$

$$b(\theta) = -log(\theta),$$

$$c(y, \phi) = \phi\, log(\phi) - log\,\Gamma(\phi) + (\phi - 1)\,log\,(y).$$

  Note: Further we use Gamma distribution for modelling of severity of specific occurances.

2.2. Link functions

There are many commonly used link functions and the choice of one of them for given purpose can be somewhat arbitrary. Although, it can be convenient to match the domain of the link function (abbreviated as link) to the support of the distribution of the response variable.

Table 2.1 Common link functions and their inverses for $\eta_i = g(\mu_i)$

Link	$g(\mu_i)$	$g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$log(\mu_i)$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$log\mu_i / (1-\mu_i)$	$1/(1+e^{-\eta_i})$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log – log	$-log(-log(\mu_i))$	$exp\left(-exp\left(-\eta_i\right)\right)$
Complementary log – log	$log(-log(1-\mu_i))$	$1 - exp\left(-exp\left(\eta_i\right)\right)$

Note: Φ is the cumulative distribution function of the standard-normal distribution.

Remind, that we are using distributions of the response variable defined as:

$$f(y, \theta, \phi) = exp\left\{\frac{y\,\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

For such density functions, which are defined with a canonical parameter θ, we introduce the *canonical link functions* g as the class of link functions for which holds:

$$g(\mu) = \theta.$$

The canonical link simplifies the GLM but other link functions may be used as well.

Table 2.2 Canonical link functions for some distributions from exponential family.

Distribution	*Canonical Link*
Normal	Identity
Binomial	Logit
Poisson	Log
Gamma	Inverse

One of the advantages of GLM is that the choice of the link function is partly separated from the distribution of the response. Although the domain of the *link function* should match to the support of the distribution of the response variable, the specific link functions, which may be used, vary from one software implementation of GLM to another. But the choice is still important and has to be done carefully. While the choice of canonical link is prefferable, neither the usage of e.g. the identity, log, inverse, or square-root links for binomial data, nor the usage of e.g. the logit, probit, log-log, or complementary log-log link for nonbinomial data are favourable.

2.3. Likelihood function

As we mentioned before, the GLM estimates are seen as the maximum likelihood estimates and so naturally we have to introduce term the likelihood function, which is crucial for the process of fitting.

Definition ([7], scalar version): The *likelihood function* of scale and canonical parameters, θ and ϕ, for a random variable Y with density function $f(y, \theta, \phi)$ is defined as

$$L_y(\theta, \phi) = f(y, \theta, \phi),$$

where y is some realization of random variable Y.

Note: For simplicity, in this section we work with scalar version of definition for *likelihood function*.

The likelihood function rewritten for some distribution from exponential family is then:

$$L_y(\theta, \phi) = \exp\left\{\frac{y.\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}.$$

Note: It is nothing else than density function of random variable Y considered as a function of parameters θ, ϕ for given y.

The *log-likelihood function* is defined as:

$$l_y(\theta, \phi) = log(L_y(\theta, \phi)) = \frac{y.\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

and by replacing the particular observation y by the random variable Y the log-likelihood function becomes random variable itself. This enables us to compute its mean value:

$$E\, l_Y(\theta, \phi) = \frac{E\, Y.\theta - b(\theta)}{a(\phi)} + E\, c(Y, \phi).$$

Important step in our search for the maximum likelihood estimate (= GLM estimate) is finding the way how to determine the value of canonical parameter θ for which the value of log-likelihood function will be in maximum. It is caused by the intivituve fact that for the most reliable estimate of response variable we consider the most probable one in terms of predetermined distribution. For such value of canonical parameter θ holds:

$$E\,\frac{\partial l_Y}{\partial \theta}(\theta, \phi) = \frac{E\, Y - b'(\theta)}{a(\phi)} = 0, \tag{1}$$

which implies the first important relationship for likelihood function and distributions from family of exponentional distributions:

$$E\, Y = b'(\theta). \tag{2}$$

Definition: The value of canonical parameter for which holds the equation (2) is called the *true value of θ.*

Now, when we know the way how could be trouhg canonical parameter the mean value estimated, we need to identify the variance of Y in order to be able to fully describe it's distribution (in most cases). From the second derivation of the likelihood function we get:

$$E\,\frac{\partial^2 l_Y}{(\partial\theta)^2} = \frac{-b''(\theta)}{a(\phi)}.$$

Theorem ([7]): For true value of θ holds:

$$E\left(\frac{\partial l_Y}{\partial \theta}\right)^2 = -E\,\frac{\partial^2 l_Y}{(\partial\theta)^2}.$$

Proof:

Let's denote support of random variable Y as $S(Y)$.

At first, we see that

$$E\left(\frac{\partial l_Y}{\partial \theta}\right) = \int_{S(Y)} \frac{\partial log(L_y(\theta, \phi))}{\partial \theta} f(y, \theta, \phi)\, dy = \int_{S(Y)} \frac{\partial log(f(y, \theta, \phi))}{\partial \theta} f(y, \theta, \phi)\, dy,$$

which after applying the chain rule

$$\frac{\partial log(f(y, \theta, \phi))}{\partial \theta} = \frac{1}{f(y, \theta, \phi)} \frac{\partial f(y, \theta, \phi)}{\partial \theta},$$

can be rewritten as

$$\int_{S(Y)} \frac{1}{f(y, \theta, \phi)} \frac{\partial f(y, \theta, \phi)}{\partial \theta} f(y, \theta, \phi)\, dy \overset{(*)}{=} \frac{\partial}{\partial \theta} \int_{S(Y)} f(y, \theta, \phi)\, dy = 0.$$

This implies

$$\int_{S(Y)} \frac{\partial log(f(y, \theta, \phi))}{\partial \theta} f(y, \theta, \phi)\, dy = 0.$$

Differentiating the last equation again by θ yields to

$$\int_{S(Y)} \frac{\partial^2 log(f(y, \theta, \phi))}{\partial \theta^2} f(y, \theta, \phi)\, dy + \int_{S(Y)} \frac{\partial log(f(y, \theta, \phi))}{\partial \theta} \frac{\partial f(y, \theta, \phi)}{\partial \theta}\, dy = 0,$$

for which after applying the same chain rule as before holds

$$\int_{S(Y)} \frac{\partial^2 log(f(y, \theta, \phi))}{\partial \theta^2} f(y, \theta, \phi)\, dy + \int_{S(Y)} \left(\frac{\partial log(f(y, \theta, \phi))}{\partial \theta}\right)^2 f(y, \theta, \phi)\, dy = 0.$$

Finally we get

$$E\left(\frac{\partial l_Y}{\partial \theta}\right)^2 = -E\frac{\partial^2 l_Y}{(\partial \theta)^2}. \qquad Q.E.D.$$

$(*)$ It is to be taken that all derivatives are evaluated at true value of θ for which is sufficient regularity. Since $f(y, \theta, \phi)$, $\frac{\partial f(y,\theta,\phi)}{\partial \theta}$ are continuous on their domains and $\int_{S(Y)} \frac{\partial f(y,\theta,\phi)}{\partial \theta}\, dy = E\frac{Y - b'(\theta)}{a(\phi)}$ for the true value of θ is equalled to 0, the order of differentiation and integration can be exchanged.

Recall the formula (1) based on which for the true value of canonical parameter θ holds:

$$E\left(\frac{\partial l_Y}{\partial \theta}\right)^2 = E\frac{Y^2 - 2\,Y(b'(\theta)) + b'(\theta)^2}{a(\phi)^2},$$

where by applying (2) we get:

$$E\left(\frac{\partial l_Y}{\partial \theta}\right)^2 = \frac{E\,Y^2 - 2\,E\,Y\,E\,Y + (E\,Y)^2}{a\,(\phi)^2} = \frac{E\,Y^2 - (E\,Y)^2}{a\,(\phi)^2} = \frac{var\,Y}{a\,(\phi)^2}.$$

It implies the second important relationship for likelihood function and distributions from family of exponentional distributions:

$$var\,Y = b''(\theta)\,a(\phi). \tag{3}$$

Recall that for purposes of this thesis we define:

$$a(\phi) = \phi / \omega$$

and hence:

$$var\,Y = \frac{b''(\theta)\,\phi}{\omega}.$$

Based on (3) we define *Variance function* V as:

$$V(b'(\theta)) = \frac{b''(\theta)}{\omega},$$

which for the true value of canonical parameter θ together with $E\,Y = \mu$ gives:

$$V(\mu) = \frac{b''(\theta)}{\omega},$$

and finally, if we use (3) and return to pointwise notation, we have:

$$var\,Y_i = V(\mu_i)\,\phi,\ i \in \{1,\ \ldots n\}, \tag{4}$$
$$E\,Y_i = b'(\theta_i).$$

Looking at the equation (4) we can see, that the variance of responses Y_i is a function of its mean μ_i and a scale parameter ϕ. This fact represents very convenient property of distributions from the exponential family.

Note: For cannonical link g, using previous equations, holds:

1. $g(\mu_i)=(b')^{-1}(\mu_i)$
2. $g'(\mu_i) = \frac{1}{V(\mu_i)}$

Table 2.3 Variance function V for some distributions from exponential family of distributions.

Distribution	*Variance function*
Normal	1
Binomial	$\frac{\mu_i.(1-\mu_i)}{n_i}$
Poisson	μ_i
Gamma	μ_i^2

Note: For binomial distribution is n_i the number of trials.

2.4. The estimation of regression coefficients

The first relationship (2) introduced in the previous section is crucial for fitting data in process of GLM. It implies that the expected mean value of Y_i, $EY_i=\mu_i$, having any distribution from family of exponentional distributions depends only on function b, which is given by distribution of Y_i and value of canonical parameter θ_i:

$$\mu_i = b'(\theta_i), \forall\ i \in \{1,\ ..., n\}.$$

This equation and the fact that from definition of the generalized linear model we know that expected mean value depends also on values of regression coefficients $\boldsymbol{\beta}$ through equation:

$$g(\mu_i) = \mathbb{X}_i\, \boldsymbol{\beta}\,, \forall\ i \in \{1,\ ..., n\},$$

leads to the equation:

$$g(b'(\theta_i)) = \mathbb{X}_i\, \boldsymbol{\beta}\,, \forall\ i \in \{1,\ ..., n\}.$$

We can see that by using likelihood function and maximimizing the value in canonical parameter θ, the estimates of regression coefficients can be computed .

In terms of GLM, instead of one random variable for the response Y, we have the disposal vector of observation $\mathbb{y} = (y_1, y_2,\ ..., y_n)$, which are realizations of vector of random variables $\mathbb{Y} = (Y_1, Y_2,\ ..., Y_n)$ with same type of distribution from the family of exponentional distributions.

Hence, it was shown in previous chapter, that for the group of functions

$$l_{y_i}(\theta_i, \phi) = log(L_{y_i}(\theta_i, \phi)) = \frac{y_i.\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi), \forall\ i \in \{1,\ ..., n\},$$

we can find corresponding group of random variables with mean values

$$E\, l_{Y_i}(\theta_i, \phi)) = E\, log(L_{Y_i}(\theta_i, \phi))) = \frac{\mu_i.\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi), \forall\ i \in \{1,\ ..., n\},$$

for which in the case of true values of cannonical parameters θ_i we have

$$b'(\theta_i) = g^{-1}(\mathbb{X}_i\, \boldsymbol{\beta}),\ \forall\ i \in \{1,\ ..., n\}.$$

We can express the canonical parameters θ_i, representing the true values of cannonical parameters, as values of some functions p_i describing relationships between θ_i and $\boldsymbol{\beta}$, via

$$p\,(\mathbb{X}_i.\boldsymbol{\beta}) := (b')^{-1}\left(g^{-1}(\mathbb{X}_i\, \boldsymbol{\beta})\right) = \theta_i,\ \forall\ i \in \{1,\ ..., n\}. \tag{5}$$

The vector $\boldsymbol{\beta}$ is supossed to be common for all response variables Y_i, which are mutually independent and have same type of distribution and so their join distribution function can be written as the product of singles ones. Hence, our goal is to determine the vector $\boldsymbol{\beta}$, through relationship (5), for which the log-likelihood function of this join

distribution for given values $\mathbb{y} = (y_1, \ldots y_n)$ is in maximum:

$$l_{\mathbb{y}}(\boldsymbol{\theta}, \phi) = \log(L_{\mathbb{y}}(\boldsymbol{\theta}, \phi)) = \log\left(\prod_{i=1}^{n} exp\left(\frac{y_i\,\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

Note: If $\mathbb{Y}$ is vector of random variables, $\mathbb{Y} = (Y_1,\ldots,Y_n)$, with vector of observations $\mathbb{y} = (y_1, \ldots, y_n)$ then:

$$l_{\mathbb{y}}(\theta, \phi) = \sum_{i=1}^{n} l_{y_i}(\theta, \phi)\,, \forall\, i \in \{1, \ldots, n\}.$$

Using the restricted form for functions a_i ($= \phi/\omega_i$), defined in section 2.1, and equation (5) the log-likelihood function can be rewritten as function of $\boldsymbol{\beta}$:

$$l_{\mathbb{y}}(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n}\left(\frac{y_i\, p\,(\mathbb{X}_i\,\boldsymbol{\beta}) - b(p\,(\mathbb{X}_i\,\boldsymbol{\beta}))}{\phi\,/\,\omega_i} + c(y_i, \phi)\right). \tag{6}$$

And now we get the formula for maximum likelihood estimate of $\boldsymbol{\beta}$, as:

$$\left(\hat{\boldsymbol{\beta}}, \hat{\phi}\right) = argmax_{\boldsymbol{\beta}\in\mathbb{R}^{k+1},\phi\in\mathbb{R}} \sum_{i=1}^{n}\left(\frac{y_i.p\,(\mathbb{X}_i\,\boldsymbol{\beta}) - b(p\,(\mathbb{X}_i\,\boldsymbol{\beta}))}{\phi\,/\,\omega_i} + c(y_i, \phi)\right),$$

where $\hat{\phi}$ is maximum likelihood estimate of scale parameter ϕ.

Hence, the maximum likelihood estimates of cannonical parameters θ_i are:

$$\hat{\theta}_i = p\left(\mathbb{X}_i.\hat{\boldsymbol{\beta}}\right) = (b')^{-1}\left(g^{-1}\left(\mathbb{X}_i.\hat{\boldsymbol{\beta}}\right)\right),\ \forall\, i \in \{1, \ldots, n\}.$$

When we are working with GLM in practise, it is useful to have an estimator which meassures the reliability of model in a similar way how *RSS* does in SLM. By enumerating the estimator in realizations $\mathbb{y}$ of vector of random variables $\mathbb{Y}$ we get an estimate called *deviance of the model* and is defined as

$$D\left(\mathbb{y}, \hat{\boldsymbol{\theta}}\right) = -2.\left(l_{\mathbb{y}}\left(\hat{\boldsymbol{\theta}}, \phi\right) - l_{\mathbb{y}}\left(\tilde{\boldsymbol{\theta}}\,, \phi\right)\right)\cdot\phi = -2\sum_{i=1}^{n}\omega_i\left(y_i\left(\hat{\theta}_i - \tilde{\theta}_i\right) - \left(b(\hat{\theta}_i) - b(\tilde{\theta}_i)\right)\right),$$

where $\tilde{\theta}_i$ indicates the estimate of cannonical parameter θ_i, $\forall\ i \in \{1, \ldots, n\}$ for the *saturated model*, the model with one parameter per observation, for which holds

$$\tilde{\mu}_i = g^{-1}\left(\mathbb{X}_i\,\tilde{\boldsymbol{\beta}}\right) = y_i,\ \forall\, i \in \{1, \ldots, n\}.$$

If the dataset is given, the value of likelihood function for the saturated model is the highest which the likelihood function could possibly have and is given by

$$\tilde{\theta}_i = (b')^{-1}\left(g^{-1}(\mathbb{y}_i)\right),\ \forall\, i \in \{1, \ldots, n\}.$$

Note: The deviance is defined to be independent from ϕ. Later we define the related

term, scaled deviance, which on the other hand depends on ϕ and is used to meassure the reliability of model.

We use deviance in next section as termination condition for Newton-Raphson algorithm. However, the main role it plays in testing of hypothesis (section 2.8.).

2.5. Maximum likelihood estimate of $\boldsymbol{\beta}$

Recall the formula (6) for maximum likelihood estimate of $\boldsymbol{\beta}$ in previous section. Now, the process of maximalization of this formula will be done by partial derivative of log-likelihood function with respect to each element of $\boldsymbol{\beta}$ and setting equal to 0. By this we get set of equations

$$\frac{\partial l_y(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{\phi / \omega_i} \left(y_i \frac{\partial p_i(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} - \frac{\partial b(p_i(\mathbb{X}_i \boldsymbol{\beta}))}{\partial \beta_j} \frac{\partial p_i(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} \right) = 0, \quad \forall\, j \in \{0,\ ...,\ k\}. \tag{7}$$

Based on formulas from previous section and by using equation (2) we have

$$\frac{\partial b(p(\mathbb{X}_i \boldsymbol{\beta}))}{\partial p(\mathbb{X}_i \boldsymbol{\beta})} = \mu_i = g^{-1}(\mathbb{X}_i \boldsymbol{\beta}),\ \forall\, i \in \{1,\ ...,\ n\},$$

which implies

$$\frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial p_i(\mathbb{X}_i \boldsymbol{\beta})} = \frac{\partial^2 b(p(\mathbb{X}_i \boldsymbol{\beta}))}{(\partial p(\mathbb{X}_i \boldsymbol{\beta}))^2},\ \forall\, i \in \{1,\ ...,\ n\}.$$

By application of the chain rule we get following formulas:

1. $$\frac{\partial b(p_i(\mathbb{X}_i.\boldsymbol{\beta}))}{\partial \beta_j} = \frac{\partial b(p_i(\mathbb{X}_i \boldsymbol{\beta}))}{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j},\ \forall\, i \in \{1,\ ...,\ n\},$$

2. $$\frac{\partial p_i(\mathbb{X}_i.\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial p_i(\mathbb{X}_i \boldsymbol{\beta})}{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j},\ \forall\, i \in \{1,\ ...,\ n\}.$$

Applying formulas 1., 2. to the set of equations (7) we get

$$\frac{\partial l_y(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\frac{\partial^2 b\,(p\,(\mathbb{X}_i \boldsymbol{\beta}))}{(\partial p\,(\mathbb{X}_i \boldsymbol{\beta}))^2}} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} \frac{1}{\phi / \omega_i} = 0,\ \forall\, j \in \{0,\ ...,\ k\}.$$

From the relationship between θ_i and $\boldsymbol{\beta}$, equation (5), we get

$$\frac{\partial l_y(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{b''(\theta_i) / \omega_i} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} \frac{1}{\phi} = 0,\ \forall\, j \in \{0,\ ...,\ k\}.$$

By using definition of variance function V, which is given by the type of distribution of the response Y_i, we get

$$\frac{\partial l_{\mathbb{y}}(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{V\left(g^{-1}(\mathbb{X}_i \boldsymbol{\beta})\right)} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} \frac{1}{\phi} = 0, \ \forall\, j \in \{0, \ ..., k\}.$$

Finally, the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ are determined via equations

$$\sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{V\left(g^{-1}(\mathbb{X}_i \boldsymbol{\beta})\right)} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} = 0, \ \forall\, j \in \{0, \ ..., k\}. \tag{8}$$

2.6. Newton-Raphson algorithm for solving non-linear equations for GLM

Coefficients $\boldsymbol{\beta}$ can be estimated by iterative approximation using iteratively reweighted least squares (IRLS). At each step of the iteration the value of likelihood function increases and by the logic of GLM the model is improved. The iteration process will end when the requested precission is achieved (the log-likelihood function does not change significantly any more). Values of $\boldsymbol{\beta}$ in the last step of iteration are considered to be the best estimates and we denote their vector by $\hat{\boldsymbol{\beta}}$.

Problem: Find $\boldsymbol{\beta}$ by solving equation (8):

$$\sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{V\left(g^{-1}(\mathbb{X}_i \boldsymbol{\beta})\right)} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} = 0, \ \forall\, j \in \{0, \ ..., k\}.$$

Solution: the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, via

Algorithm:

1. Set $p = 0$ (step counter) and $\hat{\boldsymbol{\beta}}^{[0]}$ set as in case of LSE equaled to:

$$\hat{\boldsymbol{\beta}}^{[0]} = \left(\mathbb{X}^T \mathbb{X}\right)^{-1} \mathbb{X}^T \mathbb{y}.$$

2. For current index p calculate vector of variance functions $\mathbb{V}^{[p]} = (V_1^{[p]},...,V_n^{[p]})$ as:

$$V_i^{[p]} = V\left(g^{-1}\left(\mathbb{X}_i \hat{\boldsymbol{\beta}}^{[p]}\right)\right).$$

3. Approximation of:

$$S = \sum_{i=1}^{n} \left(\frac{y_i - g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\sqrt{V_i^{[p]}}} \right)^2$$

by replacing $g^{-1}(\mathbb{X}_i.\beta)$ by its first order Taylor expansion around $\hat{\boldsymbol{\beta}}^{[p]}$ leads to an approximate value of *RSS* (pseudo *RSS*) which we are minimizing (analogically to LSE) in order to receive estimates of regression coefficients:

$$\sum_{i=1}^{n} \left(\frac{1}{\sqrt{V_i^{[p]}}} \left(y_i - g^{-1}\left(\mathbb{X}_i \hat{\boldsymbol{\beta}}^{[p]}\right) - \sum_{j=1}^{k} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} \Big|_{\hat{\boldsymbol{\beta}}^{[p]}} \left(\beta_j - \hat{\beta}_j^{[p]} \right) \right) \right)^2.$$

Pseudo data

- pseudo values of response:

$$z_i = \frac{1}{\sqrt{V_i^{[p]}}} \left(y_i - g^{-1}\left(\mathbb{X}_i \hat{\boldsymbol{\beta}}^{[p]}\right) + \sum_{j=1}^{k} \frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_j} \Big|_{\hat{\boldsymbol{\beta}}^{[p]}} \hat{\beta}_j^{[p]} \right),$$

$$\mathbb{Z} = (z_1, \ \ldots, z_n)^T.$$

- pseudo vector of predictors:

$$\mathbb{W}_i = \left(\frac{\frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_1} \Big|_{\hat{\boldsymbol{\beta}}^{[p]}}}{\sqrt{V_i^{[p]}}}, \quad \cdots \quad , \frac{\frac{\partial g^{-1}(\mathbb{X}_i \boldsymbol{\beta})}{\partial \beta_k} \Big|_{\hat{\boldsymbol{\beta}}^{[p]}}}{\sqrt{V_i^{[p]}}} \right),$$

$$\mathbb{W} = \begin{pmatrix} \mathbb{W}_1 \\ \vdots \\ \mathbb{W}_n \end{pmatrix}.$$

- pseudo *RSS*:

$$\sum_{i=1}^{n} (z_i - \mathbb{W}_i \boldsymbol{\beta})^2 = (\mathbb{Z} - \mathbb{W}\boldsymbol{\beta})^T (\mathbb{Z} - \mathbb{W}\boldsymbol{\beta}).$$

Iterative estimate $\hat{\boldsymbol{\beta}}^{[p+1]}$ is then obtained as:

$$\hat{\boldsymbol{\beta}}^{[p+1]} = arg\ min_{\beta \in \mathbb{R}^{k+1}} (\mathbb{Z} - \mathbb{W}\boldsymbol{\beta})^T (\mathbb{Z} - \mathbb{W}\boldsymbol{\beta}).$$

Solution of minimalization equals to: $\hat{\boldsymbol{\beta}}^{[p+1]} = \left(\mathbb{W}^T \mathbb{W}\right)^{-1} \mathbb{W}^T \mathbb{Z}$.

4. Check if terminating condition is met. In this work we use the condition implemented in statistical program ℝ as a part of funcion *glm*. (The practical part of this thesis is done using this program.)

Terminating condition:

$$\frac{\left| D(\mathbb{y}, \boldsymbol{\theta}^{[p+1]}) - D(\mathbb{y}, \boldsymbol{\theta}^{[p]}) \right|}{\left| D(\mathbb{y}, \boldsymbol{\theta}^{[p+1]}) \right| + 10^{-1}} < 10^{-7},$$

where $D(\mathbb{y}, \boldsymbol{\theta}^{[p+1]})$ represents deviance for the model:

$$g(E(Y_i)) = \mathbb{X}_i.\hat{\boldsymbol{\beta}}^{[p+1]}.$$

Then a vector of estimates $\boldsymbol{\theta}^{[p+1]}$ for the cannonical parameters $\boldsymbol{\theta}$ can be obtained from:

$$\boldsymbol{\theta}^{[p+1]} = \begin{pmatrix} \theta_1^{[p+1]} \\ \vdots \\ \theta_n^{[p+1]} \end{pmatrix},$$

where

$$\theta_i^{[p+1]} = p_i\left(\mathbb{X}_i \hat{\boldsymbol{\beta}}^{[p+1]}\right).$$

Note: Estimate $\phi^{[p+1]}$ of the scale parameter ϕ can be obtained from

$$\phi^{[p+1]} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{\left(Y_i - g^{-1}\left(\mathbb{X}_i \hat{\boldsymbol{\beta}}^{[p+1]}\right)\right)^2}{V_i^{[p+1]}}.$$

5. If terminating condition is not met, then increase p by one and return to the step 2, otherwise $\hat{\boldsymbol{\beta}}^{[p+1]}$ is wanted estimate $\hat{\boldsymbol{\beta}}$ and corresponding estimate $\hat{\boldsymbol{\phi}}$ represents Pearson estimate of scale parameter ϕ.

2.7. Distribution results for maximum likelihood estimates of $\boldsymbol{\beta}$

To be able to investigate the reliability of mentioned estimates we need to take a look at maximum likelihood estimators and find out which properties they have. In general, distributional results for GLM are not exact, but mostly they are based on large sample approximations, which makes use of general properties of maximum likelihood estimators including consistency. This leads to more complex study of corresponding asymptotic properties, which simplified version, based on [5], can be found in appendix together with the most important result

$$\hat{\boldsymbol{\beta}} \xrightarrow{D} N_k\left(\boldsymbol{\beta}, \boldsymbol{\mathcal{I}}^{-1}\right), \text{ where } n \to \infty \tag{9}$$

and $\boldsymbol{\mathcal{I}}$ is the Information matrix with following elements

$$\mathcal{I}_{j,l} = E\left(\frac{\partial l_Y(\boldsymbol{\beta}, \phi)}{\partial \beta_j} \frac{\partial l_Y(\boldsymbol{\beta}, \phi)}{\partial \beta_l}\right), \text{ where } j, l \in \{0, \ldots, k\}.$$

Usually the Information matrix $\mathcal{I}$ is not known and has to be estimated. Such empirical information matrix is based on theorem introduced in section 2.3. equalled to the negetive of the hessian matrix $H(-H)$, which exact formulation can be found in [4].

Note: For distributions with known scale parameter, ϕ, this result can be used to express confidential intervals directly, but if the scale parameter is unknown, e.g. gamma distribution, confidential intervals must be based on an appropriate t distribution.

2.8. Likelihood ratio tests

Now, we present tests for 2 nested models, from which we want to choose the prefferable one. There is a full model and it's submodel which omits some variables. The likelihood ratio tests indicate if the submodel's fit is significantly worse than the fit of corresponding full model.

Suppose, that we have two competitive generalized linear models:

1. $g(\mu_i) = \mathbb{X}_i \boldsymbol{\beta}$, $i \in \{1, \ldots n\}$,

2. $g(\mu_i) = \mathbb{X}_i^* \boldsymbol{\beta}^*$, $i \in \{1, \ldots n\}$,

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta}^* \in \mathbb{R}^{p^*}$, $p > p^*$ and matrix $\mathbb{X}_i^*$ is restriction of matrix $\mathbb{X}_i$. We say, that the second model represent submodel of the first one to which we adress as to full model.

Accordingly to this models we set hypothesis:

$$H_0 : \; g(\mu_i) = \mathbb{X}_i^* \boldsymbol{\beta}^* \;\; \& \;\; H_1 : g(\mu_i) = \mathbb{X}_i \boldsymbol{\beta}, \; \forall \, i \in \{1, \ldots n\}.$$

Further we assume that the scale parameter ϕ is known and $l_{\mathrm{Y}}\left(\hat{\boldsymbol{\beta}}, \phi\right)$, $l_{\mathrm{Y}}\left(\hat{\boldsymbol{\beta}}^*, \phi\right)$ are the maximized likelihoods of these models.

In such case if the hypothesis H_0 is valid, than the following holds:

$$-2\left(l_{\mathrm{Y}}\left(\hat{\boldsymbol{\beta}}^*, \phi^*\right) - l_{\mathrm{Y}}\left(\hat{\boldsymbol{\beta}}, \phi\right)\right) \xrightarrow{D} \chi^2_{p-p^*} \text{ where } n \to \infty. \tag{10}$$

Derivation of this result can be found in [6].

Note: If the hypothesis H_0 is false, then in most cases the first model has much bigger likelihood than it's submodel. Submodel has always lower likelihood than full model.

In terms of deviance, a benchmark representing full model for good fit of models is established by saturated model. Remind that the deviance, section 2.4, is defined as

$$D\left(\mathrm{y}, \hat{\boldsymbol{\theta}}\right) = -2\left(l_{\mathrm{y}}\left(\hat{\boldsymbol{\theta}}, \phi\right) - l_{\mathrm{y}}\left(\tilde{\boldsymbol{\theta}}, \phi\right)\right) \cdot \phi = -2 \sum_{i=1}^{n} \omega_i \left(y_i\left(\hat{\theta}_i - \tilde{\theta}_i\right) - \left(b_i(\hat{\theta}_i) - b_i(\tilde{\theta}_i)\right)\right),$$

where $\tilde{\theta}_i$ indicates the maximized likelihood estimate of cannonical parameter θ_i, $\forall \, i \in \{1, \ldots, n\}$ for the saturated model.

Notice, that the deviance is defined to be independent from the scale parameter, but otherwise it strongly reminds the statistic in (10). Therefore we define the related term, scaled deviance, which on the other hand depends on the estimate of scale parameter $\hat{\phi}$,

$$D^S\left(\mathrm{y}, \hat{\boldsymbol{\theta}}, \hat{\phi}\right) = \frac{D\left(\mathrm{y}, \hat{\boldsymbol{\theta}}\right)}{\hat{\phi}},$$

which can be rewritten as

$$D^S\left(\mathrm{y}, \hat{\boldsymbol{\theta}}, \hat{\phi}\right) = -2.\left(l_{\mathrm{y}}\left(\hat{\boldsymbol{\theta}}, \hat{\phi}\right) - l_{\mathrm{y}}\left(\tilde{\boldsymbol{\theta}}, \tilde{\phi}\right)\right).$$

Analysis of variance ANOVA

Further we assume that $\hat{\boldsymbol{\theta}}^*$ is maximum likelihood estimate of $\boldsymbol{\theta}$ for submodel and $\hat{\boldsymbol{\theta}}$ is maximum likelihood estimate of $\boldsymbol{\theta}$ for full model. Then for

$$D^S\left(\mathrm{y}, \hat{\boldsymbol{\theta}}^*, \phi\right) - D^S\left(\mathrm{y}, \hat{\boldsymbol{\theta}}, \phi\right) = -2.\left(l_{\mathrm{y}}\left(\hat{\boldsymbol{\theta}}^*, \phi\right) - l_{\mathrm{y}}\left(\hat{\boldsymbol{\theta}}, \phi\right)\right)$$

under H_0 holds:

- in case the parameter ϕ is known

$$D^S\left(\mathrm{y}, \hat{\boldsymbol{\theta}}^*, \phi\right) - D^S\left(\mathrm{y}, \hat{\boldsymbol{\theta}}, \phi\right) \xrightarrow{D} \chi^2_{p-p^*} \text{ when } n \to \infty.$$

- in case the parameter ϕ is unknown

$$\frac{\frac{D\left(\mathrm{y},\hat{\boldsymbol{\theta}}^*\right) - D\left(\mathrm{y},\hat{\boldsymbol{\theta}}\right)}{p-p^*}}{\frac{D\left(\mathrm{y},\hat{\boldsymbol{\theta}}\right)}{n-p}} \xrightarrow{D} F_{p-p^*,n-p} \text{ when } n \to \infty.$$

Derivation of these results can be found in [6].

Note: The advantage of the second result is that it can be used for hypothesis testing based model comparison, when ϕ is unknown. The disadvantages are the questianable assumptions about distribution of $D\left(\mathrm{y}, \hat{\boldsymbol{\theta}}\right)$ and asymptotical independence of $D\left(\mathrm{y}, \hat{\boldsymbol{\theta}}^*\right) - D\left(\mathrm{y}, \hat{\boldsymbol{\theta}}\right)$ and $D\left(\mathrm{y}, \hat{\boldsymbol{\theta}}\right)$.

3. Segemented generalized linear model

In practical usage of regression we can very often encouter cases in which the relationship between the response and the predictors changes only at certain points of the domain for some predictor X_j, $j \in \{1, \ldots, k\}$. For example, the probability of some occurance(representing the response) is strongly correlated from a certain level but up to this level predictor has no influance at all. For such cases the linear predictor(one of keystones in GLM) is very insufficient. However, as we will see, proper transformations of predictors can represent suitable solution.

In this chapter we introduce the segmented generalized linear models where the relationship between the mean of the response transformed by the link function and one or more predictors is piecewise linear. The points in which the relationship changes are called *breakpoints*. The theory justifying segmented generalized linear models can be found in [8].

The segmented generalized linear model describing the relationship between vector of response variable $\mathcal{Y} = (Y_1,\ldots,Y_n)$ and one predictor X_j with observations $x_{1,j} \ldots x_{n,j}$ and one breakpoint τ can be written as :

$$g(\mu_i) = \beta_0 + \beta_1 x_{1,j} + \beta_2 (x_{1,j} - \tau)_+ = \begin{cases} \beta_0 + \beta_1 x_{i,j} & \text{if } x_{i,j} \le \tau \\ \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{i,j} - \beta_2 \tau & \text{if } x_{i,j} > \tau \end{cases},$$
$$\forall\, i \in \{1, \ldots, n\},$$

where as ussual $\mu_i = EY_i$ and $(\cdot)_+$ is function defined as $max(\cdot, 0)$.

In general, a segmented model with m_j breakpoints $\tau(j)_1, \ldots, \tau(j)_{m_j}$ describing the relationship between the response variable $\mathcal{Y}$ and the predictors $X_{1,\ldots,}X_k$, where only one predictor X_j with obseravations $x_{1,j} \ldots x_{n,j}$ is considered to have breakpoints, is the GLM model extended by replacing j^{th} column of matrix $\mathbb{X}$ by $m_j + 1$ columns:

$$\begin{pmatrix} x_{1,j} & (x_{1,j} - \tau(j)_1)_+ & \cdots & (x_{1,j} - \tau(j)_{m_j})_+ \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,j} & (x_{n,j} - \tau(j)_1)_+ & \cdots & (x_{n,j} - \tau(j)_{m_j})_+ \end{pmatrix}.$$

<u>*Note*</u>: The reparametrization of matrix $\mathbb{X}$ can be done in analogical way for more than one predictor X_j and the general framework for the segmented generalized linear model can be achieved. From this representation it can be seen that if all breakpoints are known we get the GLM model:

$$g(\mu_i) = \mathbb{X}_i^* \boldsymbol{\beta}, \; \forall\, i \in \{1, \ldots, n\},$$

where $\mathbb{X}^*$ is matrix originated from matrix $\mathbb{X}$ with $k + 1 + \sum_{j=1}^{k} m_j$ columns.

In next two sections we demonstrate two available algorithms for segmented models with one breakpoint, which give us the basic idea how the estimations of $\boldsymbol{\beta}$ and the

unkown breakpoint τ are done and how can be generalized for more complex models.
To summarize our task, we extend GLM model

$$g(\mu_i) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_j x_{i,j} + \beta_{j+1} x_{i,j+1} + ... + \beta_k x_{i,k}\ , \forall\ i \in \{1,\ ..., n\},$$

by replacing $\beta_j x_{i,j}$ with following terms

$$\beta_{j,1} x_{i,j} + \beta_{j,2}(x_{i,j} - \tau)_+,$$

where $\beta_{j,1}$,$\beta_{j,2}$ are regression coefficients and τ is unknown breakpoint. This leads to the segmented model.

3.1. Exact algorithm

This algrorithm represents extended version of the one itroduced in [2].
We assume that the segmented generalized linear model has the following form:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{j,1} x_{i,j} + \beta_{j,2}(x_{i,j} - \tau)_+ + \beta_{j+1} x_{i,j+1} + ... + \beta_k x_{i,k},$$
$$\forall\ i \in \{1,\ ..., n\},$$

and can be rewritten as:

$$g(\mu_i) =$$
$$\beta_0 + \beta_1 x_{i,1} + ... + \beta_{j,1}\,\tau - \beta_{j,1}(\tau - x_{i,j})_+ + \beta_{j,1} x_{i,j}\left(\beta_{j,1} + \beta_{j,2}\right)(x_{i,j} - \tau)_+ + \tag{11}$$
$$\beta_{j+1} x_{i,j+1} + ... + \beta_k x_{i,k},\ \forall\ i \in \{1,\ ..., n\}.$$

By applying the following set of equations on (11):

$$\beta_0^* = \beta_0 + \beta_{j,1}\,\tau,$$
$$\beta_{j,1}^* = -\beta_{j,1},$$
$$\beta_{j,2}^* = \left(\beta_{j,1} + \beta_{j,2}\right),$$

we get

$$g(\mu_i) = \beta_0^* + \beta_1 x_{i,1} + ... + \beta_{j,1}^*(\tau - x_{i,j})_+ + \beta_{j,2}^*(x_{i,j} - \tau)_+ + \beta_{j+1} x_{i,j+1} + ... + \beta_k x_{i,k},$$
$$\forall\ i \in \{1,\ ..., n\}.$$

<u>*Note*:</u> $\beta_{j,1}^*$ is specific coefficient for the first segment ($x_{i,j} \le \tau$) and $\beta_{j,2}^*$ is for the second segment ($x_{i,j} > \tau$).

The log-likelihood function, defined in section 2.3., of such model is:

$$l_y(\boldsymbol{\beta}^*, \tau, \phi) = \sum_{i=1}^{n}\left(\frac{y_i\, p(\mathbb{X}_i^*\, \boldsymbol{\beta}^*) - b_i(p(\mathbb{X}_i^*\, \boldsymbol{\beta}^*))}{\phi / \omega_i} + c_i(y_i, \phi)\right),$$

where we keep same notation as in the case of GLM and where vector $\boldsymbol{\beta}^*$ and vectors $\mathbb{X}_i^*$ are vectors of $\boldsymbol{\beta}$ and $\mathbb{X}_i$ corresponding to the model. Again, we consider function a in it's rescricted form, ϕ/ω_i, and we are going to estimate not only the regression coefficients $\boldsymbol{\beta}^*$ but also the breakpoint τ.

Note: We add breakpoint τ as a paramater to the log-likelihood function. However, as we can see, such fucntion is not differentiable with respect to τ in $x_{i,j}$ for $\forall\, i \in \{1,\ ...,\ n\}$ and fixed $j \in \{0,\ ...,\ k\}$. Since

$$lim_{\tau \to x_{i,j+}} \frac{\partial l_{\mathbf{y}}(\boldsymbol{\beta}^*, \tau, \phi)}{\partial\, \tau} = lim_{\tau \to x_{i,j+}} \sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)}{V\left(g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)\right)} \frac{\partial g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)}{\partial \tau} \frac{1}{\phi} =$$

$$\frac{1}{\phi} \sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)\,|_{\tau = x_{i,j}}}{V\left(g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)\,|_{\tau = x_{i,j}}\right)} \left(g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)\right)'\,|_{\tau = x_{i,j}}\, \beta_{j,1}^*,$$

$$lim_{\tau \to x_{i,j-}} \frac{\partial l_{\mathbf{y}}(\boldsymbol{\beta}^*, \tau, \phi)}{\partial\, \tau} = lim_{\tau \to x_{i,j+}} \sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)}{V\left(g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)\right)} \frac{\partial g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)}{\partial \tau} \frac{1}{\phi} =$$

$$\frac{1}{\phi} \sum_{i=1}^{n} \frac{y_i - g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)\,|_{\tau = x_{i,j}}}{V\left(g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)\,|_{\tau = x_{i,j}}\right)} \left(g^{-1}(\mathbb{X}_i^*\, \boldsymbol{\beta}^*)\right)'\,|_{\tau = x_{i,j}}\, \beta_{j,2}^*$$

and so

$$lim_{\tau \to x_{i,j+}} \frac{\partial l_{\mathbf{y}}(\boldsymbol{\beta}^*, \tau, \phi)}{\partial\, \tau} \neq lim_{\tau \to x_{i,j-}} \frac{\partial l_{\mathbf{y}}(\boldsymbol{\beta}^*, \tau, \phi)}{\partial\, \tau}.$$

Hence, we can not use algorithm for solving non-linear equations based only on maximalization of log-likelihood function by setting the first derivations equaled to zero. However, the task can be divided into finite number of local maximalizations from which we can afterwards choose the global maximum likelihood estimate.

Let us reorder the observations $\left\{y_i, x_{i,1},\ ...,\ x_{i,k}\right\}_{i=1}^{n}$ with respect to the predictor X_j, i.e. $x_{i,j} \leqslant x_{i+1,j}$ for $i \in \{1,\ ...\, n-1\}$ and fixed j. So, if there is some breakpoint, the following must hold:

$$\exists\ m\ \in \{1,\ ...\, n\} : \tau = x_{m,j} \text{ or } \exists\ m\ \in \{1,\ ...\, n-1\} :\ \tau \in \left(x_{m,j},\, x_{m+1,j}\right). \tag{12}$$

In the first case, τ can be possibly equalled to $x_{i,j}$ for some $i \in \{1,\ ...,\ n\}$ and so we have at most n different sets of log-likelihood nonlinear equations:

$$l_{\mathbf{y}}(\boldsymbol{\beta}^*, x_{i,j}, \phi) = \sum_{i=1}^{n} \left(\frac{y_i . p_i\,(\mathbb{X}_i^*\, \boldsymbol{\beta}^*) - b_i(p_{\ i}\,(\mathbb{X}_i^*\, \boldsymbol{\beta}^*))}{\phi\, /\, \omega_i} + c_i(y_i, \phi) \right),$$

which can be solved by same Newton-Raphson algorithm as was introduced in section 2.6. Resolving this equations we get possibly n different values of log-likelihood function, from which we choose the maximal one and the correspondig value $x_{.,j}$ is considered to be the best estimate of breakpoint τ, $\hat{\tau}$. We also get the estimate of $\boldsymbol{\beta}^*$.

However, the estimate from the second part of alternative (12) can have even higher value of likelihood function and so we also have to consider the case $\tau \in (x_{m,j}, x_{m+1,j})$. We get possibly $n-1$ different sets of log-likelihood nonlinear equations, where we assume that the unknown breakpoint is between $(x_{m,j}, x_{m+1,j})$ for some $m\ \in \{1,\ ...\, n-1\}$.

In this case, the corresponding model can be divided into two parts:

- For $i \in \{1, \ ..., m\}$:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i,1} + ... - \beta^*_{j,1} x_{i,j} + \beta_{j+1} x_{i,j+1} + ... + \beta_k x_{i,k}$$

- For $i \in \{m+1, \ ..., n\}$:

$$g(\mu_i) = \beta^*_0 + \beta_1 x_{i,1} + ... + \beta^*_{j,2}(x_{i,j} - \tau) + \beta_{j+1} x_{i,j+1} + ... + \beta_k x_{i,k} ,$$

$$g(\mu_i) = \beta^{**}_0 + \beta_1 x_{i,1} + ... + \beta^*_{j,2} x_{i,j} + \beta_{j+1} x_{i,j+1} + ... + \beta_k x_{i,k} ,$$

where $\beta^{**}_0 = \beta^*_0 - \beta^*_{j,2} \tau$.

And so matrix form of model is:

$$g(\boldsymbol{\mu}) = \begin{pmatrix} 1 & 0 & x_{1,1} & \cdots & x_{1,j-1} & x_{1,j} & 0 & x_{1,j+1} & \cdots & x_{1,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \vdots & \ddots & \vdots & x_{m,j} & 0 & \vdots & \ddots & \vdots \\ 0 & 1 & \vdots & \ddots & \vdots & 0 & x_{m+1,j} & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & x_{n,1} & \cdots & x_{n,j-1} & 0 & x_{n,j} & x_{n,j+1} & \cdots & x_{n,k} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta^{**}_0 \\ \beta_1 \\ \vdots \\ -\beta^*_{j,1} \\ \beta^*_{j,2} \\ \beta_{j+1} \\ \vdots \\ \beta_k \end{pmatrix}.$$

This model already can be solved by Newton-Raphson algorithm introduced in 2.6.

Recall that:

$$\beta^*_0 = \beta_0 - \beta^*_{j,1} \tau,$$

$$\beta^{**}_0 = \beta^*_0 - \beta^*_{j,2} \tau,$$

what implies:

$$\hat{\tau} = \frac{\hat{\beta}_0 - \hat{\beta}^{**}_0}{\hat{\beta}^*_{j,1} + \hat{\beta}^*_{j,2}}.$$

If $\hat{\tau} \in (x_{m,j}, x_{m+1,j})$, we have found a local maximum, otherwise there is no local maximum within interval $(x_{m,j}, x_{m+1,j})$. We repeat following process for every $m \in \{1, \ ..., n-1\}$. At the end we have possibly $n-1$ different values of local maximum for intervals $(x_{m,j}, x_{m+1,j})$.

Finally, the global maximum of log-likelihood function, and corresponding estimate of regression coeficients $\boldsymbol{\beta}^*$ and breakpoint $\hat{\tau}$ is given by the maximum of this finite number of local maxima.

This algorithm gives us very good idea how the breakpoints can be estimated, but because of its computational complexity it is not very efficient. Also, the corresponding algorithms for extensions to more breakpoints become very complicated.

3.2. Algorithm based on approximate linear representation

The idea comes from possibility to approximate the non-linear term:

$$(x_{i,j} - \tau)_+$$

by a first-order Taylor expansion around an initial known value $\tau^{(0)}$:

$$\left(x_{i,j} - \tau^{(0)}\right)_+ + \left(\tau^{(0)} - \tau\right).I_{\left(x_{i,j} > \tau^{(0)}\right)},$$

where $I_{\left(x_{i,j} > \tau^{(0)}\right)} = \begin{cases} 0 & x_{i,j} \le \tau^{(0)} \\ 1 & x_{i,j} > \tau^{(0)} \end{cases}$.

Note, that:

$$\frac{\partial (x_{i,j} - \tau)_+}{\partial \tau} \Big|_{\tau^{(0)}} = -I_{\left(x_{i,j} > \tau^{(0)}\right)}.$$

In the first step of algorithm we set the step counter p to 0 and choose the initial value of breakpoint $\tau^{(0)}$. The initial value $\tau^{(0)}$ can represent our expert guess of breakpoint value. The p^{th} step of algorithm starts by stating the standardized model with breakpoint $\tau^{(p)}$

$$g(\mu_i) = \beta_0 + ... + \beta_{j,1}\, x_{i,j} + \beta_{j,2}.\left(\left(x_{i,j} - \tau^{(p)}\right)_+ + \left(\tau^{(p)} - \tau\right) I_{\left(x_{i,j} > \tau^{(p)}\right)}\right) + ... + \beta_k\, x_{i,k},$$
$$\forall\, i \in \{1,\ ..., n\},$$

which is reparametrized as

$$g(\mu_i) = \beta_0 + ... + \beta_{j,1}\, x_{i,j} + \beta_{j,2}.\left(x_{i,j} - \tau^{(p)}\right)_+ + \beta_{j,3}\, I_{\left(x_{i,j} > \tau^{(p)}\right)} + ... + \beta_k\, x_{i,k},$$
$$\forall\, i \in \{1,\ ..., n\},$$

using

$$\beta_{j,3} = \beta_{j,2}\left(\tau^{(p)} - \tau\right).$$

Now, for this parametrized model we can use Newton-Raphson algorithm introduced in the section 2.6. and receive estimates $\hat{\beta}_{j,2}, \hat{\beta}_{j,3}$, from which we get estimate $\hat{\tau}$:

$$\hat{\tau} = \frac{\hat{\beta}_{j,2}\, \tau^{(p)} - \hat{\beta}_{j,3}}{\hat{\beta}_{j,2}}.$$

If the absolute value of estimate $\hat{\beta}_{j,3}$ is lower than our predetermined terminate condition, the algorithm stops and there is no significant improvement for $\hat{\tau}$ in terms of breakpoint estimate. Otherwise the counter p increases by 1 and the cycle repeats.

If the breakpoint exists, the algorithm in a deterministic model converges and therefore $\hat{\tau}$ is assumed to be maximum likelihood estimate. This algorithm may be easily extended for models with more breakpoints only by including the appropriate constucted variables for additional breakpoints. So, at each step the estimates of all breakpoints are updated via cycle of algorithm and this is a very efficient way how to perform muliple breakpoint estimation.

Note: This algorithm is part of package *segmented* implemented in statistical program ℝ and is used in practical part of this thesis.

4. Generalized Additive Model (GAM)

In this chapter we demonstrate relationships in which GLM would not be able to deliver precise fit. It is caused by the fact, that there are missing effective tools, besides segmentation for reparametrization of predictors. Hence, we introduce Generalized Additive Model (GAM), which extend possibilities of GLM by using splines as reparametrization tool. Splines are piecewise polynomias characterized as smooth (of class C^{∞}) functions defined over predictors. In this thesis we introduce one-dimensional regression splines (over one predictor) and two-dimensional regression splines (over a pair of predictors). We focus on penalized regression splines, which represent middle way between regression splines and smoothing splines.

At this place we would like to give some informal definitions([1]) related to regression splines.

Regression spline is a non-parametric regression technique, which models non-linearities and interactions between variables. The data are fitted to a set of spline basis functions (see 4.1.). On the other hand, *smooting spline* is a parametric method used to fit a smooth curve to a set of given observations. There is a *smoothing parameter* λ, which has to controll the balance between good fitting of data and overfitting the data. The very importatnt part of this method is *roughness penalty,* indicating how much is the curve overfitting. Finally, *penalized regression splines* is method using the spline basis functions and also the penalty typicall for smoothnig splines.

We assume that the proper form of model is

Generalized Additive Model with one and two-dimensiol splines

$$g(\mu_i) = \beta_0 + f_{1,1}(x_{i,1}, x_{i,1}) + \ldots + f_{1,k}(x_{i,1}, x_{i,k}) + f_{2,2}(x_{i,2}, x_{i,2}) + \ldots + f_{k,k}(x_{i,k}, x_{i,k}),$$
$$i \in \{1, \ldots, n\},$$

where :

- **i.** For all variables the assumptions of GLM are valid.
- **ii.** $f_{j,j}(x_{i,j}, x_{i,j}) = f_j(x_{i,j})$ for $\forall\, j \in \{1, \ldots, k\}$, $i \in \{1, \ldots, n\}$ are one-dimensional splines.
- **iii.** $f_{j,l}(x_{i,j}, x_{i,l})$ for $\forall\, j, l \in \{1, \ldots, k\}$, $j \neq l$ and $\forall\, i \in \{1, \ldots, n\}$ are two dimensional splines.

<u>*Note*:</u> For following choices of spline functions:

$$f_{j,j}(x_{i,j}, x_{i,j}) = x_{i,j}, \quad \forall\, j \in \{1, \ldots, k\}, \forall\, i \in \{1, \ldots, n\},$$
$$f_{j,l}(x_{i,j}, x_{i,l}) = 0, \quad \forall\, j, l \in \{1, \ldots, k\}, \; j \neq l, \forall\, i \in \{1, \ldots, n\}$$

we get the form of GLM as was introduced in the second chapter.

In GAM we have to specify a basis for penalized regression splines, along with a corresponding definition of what is meant by the smoothness of spline (sections 4.3, 4.4). And while correct choices of basis and smoothing parameters allow us to achieve very precise fit, on the other hand poor choices could lead to the unreliable estimates or even to overfitting the data. This brings big amount of subjectivity to the modelling and so the process requires besides statistical knowledges also deep understanding of dataset.

Note: In section 4.4, we see the main result of this chapter, namely that by choice of the bases and roughness penalty we turn GAM into penalized GLM with regression coefficients $\boldsymbol{\beta}$ and smoothing parameters $\boldsymbol{\lambda}$.

Firstly we have to choose a set of basis functions for each spline, so that:

1. One-dimenional regression splines could be represented by regression B-splines or regression thin plate splines.
2. Two-dimenional regression splines could be represented by regression thin plate splines.

4.1. Regression B-splines for one-dimensional splines

The regression B-splines (abbreviated from Basis splines) are constructed piecewise from polynomial functions via

$$f_j(x) = \sum_{h=1}^{m_j} \beta(j)_h \, b(j)_h \, (x),$$

where:

- **i.** x belongs to the support of X_j,
- **ii.** $\beta(j)_h$ are regression coefficients, $h = 1, \ ..., \ m_j$,
- **iii.** $b(j)_h \, (x)$ are polynoms known as basis functions.

Note: We use $\beta(j)_h$ as notation for $\beta_{j,h}$ due to bigger transparency in the following sections.

And so we assume that the effect of predictor X_j for some $j \in \{1,...,k\}$ can be approximated by a polynomial spline written in terms of linear combination of basis functions. Degree of spline f_j, d_j, is defined as maximal degree of polynoms $\{b(j)_1, b(j)_2, \ ..., \ b(j)_{m_j}\}$. Corresponding theory to the following results can be found in [3].

As was introduced in previous chapter, the relationship between response and predictors can change in some values of predictors. We dealt with this problem by using breakpoints, which could be automatically determined in process of estimation.

In terms of GAM is with this changes in behaviour dealt by adding *knots*. Set of knots represents values of predictors, in which the splines are constructed. The choice of knot's positions is very important and has big impact on estimation. Hence, the choice of number and position of knots becomes a crucial problem.

In terms of B-Splines for the positions of knots we can use preliminary estmimates. Such as:

1. The knots can be selected manually, what is common in practise and for those, who understand the dataset very well can be even prefferable.
2. The knots are selected such that they segment the values of predictors into groups of equal size.
3. The knots are distributed at equal distances between the minumum and the maximum values of predictors.
4. The knots are selected by random sampling.

Let $\boldsymbol{t}(j) = \{t(j)_0, t(j)_1, \ ..., t(j)_{p_j}\}$ be a sequence of such preliminary estimates of knot positions for the predictor X_j, for which holds:

$$t(j)_0 = \min\{x_{1,j}, \ ..., x_{n,j}\},$$
$$t(j)_{p_j} = \max\{x_{1,j}, \ ..., x_{n,j}\},$$
$$t(j)_0 \leq t(j)_1 \leq ... \leq t(j)_{p_j}.$$

To such sequences $\boldsymbol{t}(j)$ for $j \in \{1, \ .., k\}$ we from now on refer as to the *knot sequences*. Small number of knots can cause that the spline is not flexible enough to capture the effect of predictor but on the other hand a big number of knots can lead to overfitting. The balance can be achieved by adding the roughness penalty(section 4.3), which should prevent from overfitting.

For the number of knots , p_j, must hold

$$p_j > d_j + h, \ \forall \ j \in \{1, \ ..., k\} \tag{13}$$

where d_j is equalled to degree of f_j and h is index of the knot from sequence $\boldsymbol{t}(j)$.

For each of the knots $t(j)_h$, where for the index h the inequality (13) holds, of sequence $\boldsymbol{t}(j)$ for the predictor X_j we introduce a *h-th* B-splines basis function $b(j)_h$, where the order of the B-spline basis equals d_j.

- Definition of *h-th B-spline basis of order* d_j for spline f_j is given recursively on the degree l of basis funcion $b(j)_h$.
  - $l = 0$

$$b(j)_h^{[0]}(x) = \begin{cases} 1 \text{ for } x \in \langle t(j)_h, t(j)_{h+1}) \\ 0 \text{ otherwise} \end{cases}.$$

- $l \geq 1$

$$b(j)_h^{[l]}(x) = u(j)_h^{[l]}(x)\, b(j)_h^{[l-1]}(x) + \left(1 - u(j)_{h+1}^{[l]}(x)\right) b(j)_{h+1}^{[l-1]}(x),$$

where:

$$u(j)_h^{[l]}(x) = \begin{cases} \frac{x - t(j)_h}{t(j)_{h+l} - t(j)_h} & \text{if } t(j)_h \neq t(j)_{h+l} \\ 0 & \text{otherwise} \end{cases}.$$

And so, for $l = d_j$ we can introduce the *h-th* B-spline basis function of order d_j as:

$$b(j)_h(x) = b(j)_h^{[d_j]}(x).$$

A B-spline f_j of degree d_j is defined as a linear combination of $m_j = p_j - d_j - 1$ B-spline basis functions:

$$f_j(x) = \sum_{h=1}^{m_j} \beta(j)_h\, b(j)_h(x).$$

Obviously, the generality of this process suffers by subjective choice of the position of knots. However, there are splines introduced in next section which deals with knots in more general way.

4.2. Thin plate splines for one- and two-dimensional splines

Thin plate splines (TPS) are very elegant and general solution to the problem of estimating a regression spline for multilpe predictors. They are considered to be an ideal "smoother" , because they were designed to reach exact agreement between smoothness and fitting data. Theoretical basis are given in [4]. The general design of TPS even solve crucial problem of choice of knot positions because basis functions are emerged according to observed data.

In this thesis we are using the general form of regression TPS, which can be found in [6]:

- One-dimensional regression thin plate spline for predictor X_j is defined as

$$f_j(x) = \beta(j)_0\, x + \sum_{i=1}^{n} \beta(j)_i\, \eta(j)_i(x),$$

where

i. n is number of observations,

ii. x belongs to the support of X_j,

iii. $\beta(j)_i$ for $i \in \{0, \ldots, n\}$ are regression coefficients,

iv. x and $\eta(j)_i(x) = \frac{1}{12} |x_{i,j} - x|^3$ for $i \in \{1, \ldots, n\}$ are basis functions.

- Two-dimensional regression thin plate spline for predictors X_j and X_l for $j \neq l$ is defined as:

$$f_{j,l}(x, y) = \beta(j, l)_0^1 x + \beta(j, l)_0^2 y + \sum_{i=1}^{n} \beta(j, l)_i \eta(j, l)_i (x, y),$$

where

i. n is number of observations,

ii. x belongs to the support of X_j, y belongs to the support of X_l,

iii. $\beta(j, l)_0^1$,$\beta(j, l)_0^2$,$\beta(j, l)_i$ for $i \in \{1, \ ..., n\}$ are regression coefficients,

iv. x, y and

$$\text{for } (x - x_{i,j})^2 + \left(y - x_{i,l}\right)^2 \neq 0:$$
$$\eta(j, l)_i (x, y) = \frac{1}{16\pi} \left((x - x_{i,j})^2 + \left(y - x_{i,l}\right)^2\right) log\left((x - x_{i,j})^2 + \left(y - x_{i,l}\right)^2\right)$$
$$\text{for } (x - x_{i,j})^2 + \left(y - x_{i,l}\right)^2 = 0: \eta(i, j)_i (x, y) = 0$$

for $i \in \{1, \ ..., n\}$ are basis functions.

On the other hand, when general form of TPS is used, computational costs can be large. If the dataset is large, number of parameters also goes up and the computational complexity for estimation of the model fit is proportional to the cube of the number of parameters. Such computional costs are very high price to pay for usage of these splines. Effective number of needed spline paramaters is in fact usually a small proportion of dataset range. Hence, it seems wasteful to use so many parameters to represent the model. This brings the question, whether approximation of TSP could be produced with lower computational complexity.

And in the most cases of practical usage as well as in the practical part of this thesis is used the regression TSP based on preleminary estimates of knot positions similar to the previous section.

- One-dimensional regression thin plate spline for predictor X_j with knots $t(j)_1, \ ..., t(j)_{p_j}$ is defined as:

$$f_j(x) = \beta(j)_0 x + \sum_{h=1}^{p_j} \beta(j)_h \eta(j)_h (x),$$

where

i. x belongs to the support of X_j,

ii. $t(j)_h$ for $h \in \{1, \ ..., p_j\}$ are preliminary estimates of knot positions for X_j,

iii. $\beta(j)_h$ for $h \in \{0, \ ..., p_j\}$ are regression coefficients,

iv. x and $\eta(j)_h (x) = \frac{1}{12} |t(j)_h - x|^3$ for $h \in \{1, \ ..., p_j\}$ are basis functions.

- Two-dimensional regression thin plate spline for predictors X_j and X_l for $j \neq l$ with pairs of knots $\left[t(j, l)_1^1, t(j, l)_1^2\right]$, ..., $\left[t(j, l)_{p_{j,l}}^1, t(j, l)_{p_{j,l}}^2\right]$ is defined as:

$$f_{j,l}(x, y) = \beta(j, l)_0^1 x + \beta(j, l)_0^2 y + \sum_{h=1}^{p_{j,l}} \beta(j, l)_h \eta(j, l)_h (x, y),$$

where

i. x belongs to the support of X_j, y belongs to the support of X_l,

ii. $t(j, l)_h^1$ for $h \in \{1, ..., p_{j,l}\}$ are preliminary estimates of knot positions for X_j, $t(j, l)_h^2$ for $h \in \{1, ..., p_{j,l}\}$ are preliminary estimates of knot positions for X_l,

iii. $\beta(j, l)_0^1, \beta(j, l)_0^2, \beta(j, l)_h$ for $h \in \{1, ..., p_{j,l}\}$ are regression coefficients,

iv. x, y and

for $\left(x - t(j, l)_h^1\right)^2 + \left(y - t(j, l)_h^2\right)^2 \neq 0$:

$$\eta(j, l)_h (x, y) = \frac{1}{16\pi} \left(\left(x - t(j, l)_h^1\right)^2 + \left(y - t(j, l)_h^2\right)^2\right) log\left(\left(x - t(j, l)_h^1\right)^2 + \left(y - t(j, l)_h^2\right)^2\right)$$

for $\left(x - t(j, l)_h^1\right)^2 + \left(y - t(j, l)_h^2\right)^2 = 0$: $\eta(j, l)_h (x, y) = 0$

for $h \in \{1, ..., p_{j,l}\}$ are basis functions.

For conciseness we introduce notatation:

1. $f_j \equiv f_j(x)$, $f_{j,l} \equiv f_{j,l}(x, y)$,
2. $\boldsymbol{\beta}(j)$ - vector of regression coefficients of one-dimensional spline f_j for predictor X_j,
3. $\boldsymbol{\beta}(j, l)$- vector of regression coefficients of two-dimensional spline $f_{j,l}$ for predictors X_j,X_l.

Note: Spline basis functions can be used as transformations of predictors without penalties (sections 4.3,4.4) and such models can be seen as GLM. To this approach we reffer as to the *simple regression spline* approach. In a simple regression spline approach the regression coefficients can be estimated using same algorithm as was introduced in section 2.6. However the risk of overfitting the data is very high. To overcome the problem of finding the balance between overfitting and ensuring sufficient level of fitting we use the penalized likelihood estimation, which we introduce in next sections.

4.3. Roughness penalty and regression splines

In this thesis we control the model's smoothness by adding penalty for too "twisting shape" of splines. This penalty is defined

- For one-dimensional spline for predictor X_j:

$$J(f_j) = \int_{S(X_j)} \left(\frac{\partial^2 f_j(x)}{\partial x^2} \right)^2 dx,$$

- For two-dimensional spline for predictors X_j and X_l, $j \neq l$:

$$J(f_{j,l}) = \int_{S(X_j)} \int_{S(X_l)} \left(\frac{\partial^2 f_{j,l}(x, y)}{\partial x^2} \right)^2 + \left(\frac{\partial^2 f_{j,l}(x, y)}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f_{j,l}(x, y)}{\partial y^2} \right)^2 dx\, dy,$$

where

i. $S(X_j)$ is support of X_j and $S(X_l)$ is support of X_l,

ii. x belong to $S(X_j)$, y belong to $S(X_l)$,

iii. f_j and $f_{j,l}$ are corresponding spline functions.

The integrated square of second derivative of spline function penalizes models which include too "twisty" splines. On the other hand with these "twisty" splines can be achieved better fits of data. The trade-off between model fit and model smoothness is then controlled by smoothing parameters λ_j, $\lambda_{i,l}$. Theory justifying this form of penalty can be found in [6].

Note: Penalty based on the second derivative is the most common form of penalty used in modern statistics although the method can easily be adapted to penalties based on other derivatives.

Because $f_j, f_{j,l}$ are linear in the regression coefficients $\boldsymbol{\beta}(j)$, $\boldsymbol{\beta}(j, l)$, the penalties can be written as:

- $J(f_j) = \boldsymbol{\beta}(j)^T . \mathbb{S}(j) . \boldsymbol{\beta}(j)$,
  where $\boldsymbol{\beta}(j)$ is vector of dimension $p_j + 1$ and $\mathbb{S}(j)$ is matrix $(p_j + 1) \times (p_j + 1)$,
- $J(f_{j,l}) = \boldsymbol{\beta}(j, l)^T . \mathbb{S}(j, l) . \boldsymbol{\beta}(j, l)$, where β(j,l) is vector of dimension $p_{j,l} + 2$ and $\mathbb{S}(j, l)$ is matrix $(p_{j,l} + 2) \times (p_{j,l} + 2)$.

Recall that p_j and $p_{j,l}$ are numbers of knots for corresponding splines. Thus, $\mathbb{S}(j)$ and $\mathbb{S}(j, l)$ are matrices of known values, which depend only on basis of particular splines. For more details we recommend [7].

4.4. Turning GAM into penalized GLM

For the general form of GAM (used in this thesis)

$$g(\mu_i) = \beta_0 + f_1(x_{i,1}) + ... + f_{1,k}(x_{i,1}, x_{i,k}) + f_2(x_{i,2}) + ... + f_{k-1,k}(x_{i,k-1}, x_{i,k}) + f_{k,k}(x_{i,k}, x_{i,k}), \forall\, i \in \{1, ..., n\},$$

is the *penalized likelihood function* defined as:

$$l_y^P(\boldsymbol{\beta}, \boldsymbol{\lambda}, \phi) = l_y(\boldsymbol{\beta}, \phi) - \sum_{j=1}^{k} \frac{\lambda_j}{2} \boldsymbol{\beta}(j)^T \mathbb{S}(j) \boldsymbol{\beta}(j) - \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} \frac{\lambda_{j,l}}{2} \boldsymbol{\beta}(j, l)^T \mathbb{S}(j, l) \boldsymbol{\beta}(j, l), \qquad (14)$$

where:

- **i.** $\boldsymbol{\beta} = (\boldsymbol{\beta}(1)^T, \boldsymbol{\beta}(1, 2)^T, ..., \boldsymbol{\beta}(k)^T)$,
- **ii.** $l_y(\boldsymbol{\beta}, \phi)$ is the log-likelihood function for simple regression aproach recall note in section 4.2.,
- **iii.** $\lambda_j, \lambda_{j,l}$ are the smoothing parameters and $\mathbb{S}(j)$, $\mathbb{S}(j, l)$ are the matrices defined above.

Smoothing parameters determining the level of trade-off between fitting the data and smoothness of splines. The higher the λ_j, $\lambda_{j,l}$ are, the smoother the estimated splines $\hat{f}_j, \hat{f}_{j,l}$ will be. But if they are too high then the splines probably overfit the data and if they are too low then the splines need not to fit the data on sufficient level.

The GAM fitting objective (14) can be defined in terms of the model deviance as:

$$\hat{\boldsymbol{\beta}} = arg\,min_{\boldsymbol{\beta}} \left(D(\mathbb{y}, \boldsymbol{\beta}) + \sum_{j=1}^{k} \lambda_j \boldsymbol{\beta}(j)^T \mathbb{S}(j) \boldsymbol{\beta}(j) + \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} \lambda_{j,l} \boldsymbol{\beta}(j, l)^T \mathbb{S}(j, l) \boldsymbol{\beta}(j, l) \right), \qquad (15)$$

where $D(\mathbb{y}, \boldsymbol{\beta})$ is deviance defined in section 2.4. written in terms of regression coefficients $\boldsymbol{\beta}$ based on (5).

For given smoothing parameters can be this objective quadratically approximated by

$$\hat{\boldsymbol{\beta}} = arg\,min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} w_i (z_i - \mathbb{X}_i \boldsymbol{\beta})^2 + \sum_{j=1}^{k} \lambda_j \boldsymbol{\beta}(j)^T \mathbb{S}(j) \boldsymbol{\beta}(j) + \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} \lambda_{j,l} \boldsymbol{\beta}(j, l)^T \mathbb{S}(j, l) \boldsymbol{\beta}(j, l) \right),$$

where w_i, z_i for $\forall\, i \in \{1, ..., n\}$ are original Newton based version of pseudodata which can be found in [6].

Such approximation should reasonably capture the dependence of the penalized deviance on the smoothing parameters and $\boldsymbol{\beta}$, in the vicinity of the current choices of the smoothing parameters, and the corresponding minimizing values of $\boldsymbol{\beta}$. Penalized likelihood maximization can only estimate model coefficients, $\boldsymbol{\beta}$, if the smoothing parameters are given.

Before covering the estimation of smoothing pamarameters, we have to introduce number of degrees of freedom for GAM. Notice if the smoothing parameters were all set to zero then we would get (unconstrained) GLM and so number of the degrees of

freedom would be the dimension of $\boldsymbol{\beta}$. On the other hand, if all the smoothing parameters are very high then the model is quite inflexible and therefore number of degrees of freedom would be very low.

Let define following terms

$$\mathbb{S} = \sum_{j=1}^{k} \lambda_j \mathbb{S}(j) + \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} \lambda_{j,l} \mathbb{S}(j, l),$$

$$\mathbb{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \cdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_n \end{pmatrix}$$

The matrix of effective degrees of freedom becomes

$$\mathbb{F} = \left(\mathbb{X}^T \mathbb{W}\mathbb{X} + \mathbb{S}\right)^{-1} \mathbb{X}^T \mathbb{W}\mathbb{X}$$

Then the number of effective degrees of freedom (EDF) is tr($\mathbb{F}$). Effective degrees of freedom for individual smooths are found by summing the corresponding $\mathbb{F}_{j,j}$ values for their coefficients.

Smoothing parameters can be estimated for example via *cross-validation*. The idea of this method is to delete one observation at a time from dataset and try to predict it from the model fitting to remaining observations.

Ordinary cross-validation score is defined as :

$$CV(\boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} D^{[-i]}\left(\mathbb{y}, \hat{\boldsymbol{\beta}}\right)$$

This score results from leaving out one observation (leaving i^{th} observation is denoted by superscript $[-i]$) in each turn, fitting the model to the remaining data and calculating it's deviance $D^{[-i]}\left(\mathbb{y}, \hat{\boldsymbol{\beta}}\right)$ on the whole dataset. The goal is to choose a smoothing parameter, that produces the best predictions on data, which haven't been analyzed and so we choose the one which minimize CV($\boldsymbol{\lambda}$). Unfortunately, this is too inefficient and it is needed a big computional complexity to calculate CV($\boldsymbol{\lambda}$) by leaving out one observation at a time and fitting the model to each of the n remaining observations. In practise are more often used:

- in case that scale parameter ϕ is given, *UBRE score* defined as :

$$UBRE(\boldsymbol{\lambda}) = D\left(\mathbb{y}, \hat{\boldsymbol{\beta}}\right) + 2\,\phi\,\mathrm{tr}(\mathbb{F}),$$

- in case that scale parameter ϕ is not given, *General cross validation score* defined as:

$$GCV(\boldsymbol{\lambda}) = \frac{n^{-1}\, D\left(\mathrm{y}, \hat{\boldsymbol{\beta}}\right)}{\left(1 - n^{-1}\,\mathrm{tr}(\mathbb{F})\right)^2}.$$

Substantiations and specifics for these and following results can be found in [7].

To sum it up, the maximum for penalized likelihood function can be achieved by penalized *Iteratively Reweighted Least Squares* (IRLS) for fitting objective (15). When at each iteration a penalized weighted least squares problem is solved and the smoothing parameters of that problem are estimated by GCV/UBRE. Eventually, both regression coefficients and smoothing parameter estimates converge.

It can be shown, also in [7], that solution of this problem has the form:

$$\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^T \mathbb{W}\, \mathbb{X} + \mathbb{S}\right)^{-1} \mathbb{X}^T \mathbb{W}\grave{}\mathrm{y}, \text{ where } n \to \infty.$$

And the covariance matrix for the estimators $\hat{\boldsymbol{\beta}}$ is equaled to

$$\mathbb{V}_{\hat{\boldsymbol{\beta}}} = \left(\mathbb{X}^T \mathbb{W}\, \mathbb{X} + \mathbb{S}\right)^{-1} \mathbb{X}^T \mathbb{W}\, \mathbb{X} \left(\mathbb{X}^T \mathbb{W}\, \mathbb{X} + \mathbb{S}\right)^{-1} \phi.$$

The scale parameter can be estimated by

$$\hat{\phi} = \frac{\sum_{i=1}^{n} w_i \left(z_i - \mathbb{X}_i\, \hat{\boldsymbol{\beta}}\right)^2}{n - \mathrm{tr}(\mathbb{F})}.$$

And as distributional result we have

$$\hat{\boldsymbol{\beta}} \xrightarrow{D} N_{\dim(\boldsymbol{\beta})}\left(E(\hat{\boldsymbol{\beta}}), \mathbb{V}_{\hat{\boldsymbol{\beta}}}\right), \text{ where } n \to \infty$$

<u>*Note*:</u> Generally, it does not hold $E(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$. However, if $\boldsymbol{\beta}=\mathbf{0}$ then $E(\hat{\boldsymbol{\beta}})=\mathbf{0}$, with the same validity approximately for some subsets of $\boldsymbol{\beta}$. Therefore, this result can be used for testing significance of regression coefficients.

Now we introduce distributional results used by Wald tests for expressing the corresponding *p-values*.

We state the null hypothesis for some subvector of $\boldsymbol{\beta}$,$\boldsymbol{\beta}_j$, with dimension equaled to d:

$$H_0 : \boldsymbol{\beta}_j = \mathbf{0}$$

Under this null hypothesis following claims hold:

Based on previous note we get

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} \xrightarrow{D} N_{\dim(\beta_j)}\left(0, \mathbb{V}_{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}}\right), \text{ where } n \to \infty.$$

Therefore, if $\mathbb{V}_{\hat{\boldsymbol{\beta}}_j}$ is of full rank we can use

$$\hat{\boldsymbol{\beta}}_j^{\boldsymbol{T}} \mathbb{V}_{\hat{\boldsymbol{\beta}}_j}^{-1} \hat{\boldsymbol{\beta}}_j \xrightarrow{D} \chi^2_d,$$

where $n \to \infty$ and d is dimension of $\boldsymbol{\beta}_j$.

Note: Penalization usually causes that the covariance matrices $\mathbb{V}_{\hat{\boldsymbol{\beta}}_j}$ are very often not of full rank.

If $\mathbb{V}_{\hat{\boldsymbol{\beta}}_j}$ is of rank $r < d$ we define $\mathbb{V}^{r-}_{\hat{\boldsymbol{\beta}}_j}$ as the r-rank pseudoinverse of the covariance matrix $\mathbb{V}_{\hat{\boldsymbol{\beta}}_j}$ and we can use

$$\hat{\boldsymbol{\beta}}_j^{\boldsymbol{T}} \mathbb{V}^{r-}_{\hat{\boldsymbol{\beta}}_j} \hat{\boldsymbol{\beta}}_j \xrightarrow{D} \chi^2_r, \text{ where } n \to \infty.$$

If $\mathbb{V}_{\hat{\boldsymbol{\beta}}_j}$ contains an unknown scale parameter then we can use

$$\frac{\dfrac{\hat{\boldsymbol{\beta}}_j^{\boldsymbol{T}} \mathbb{V}^{r-}_{\hat{\boldsymbol{\beta}}_j} \hat{\boldsymbol{\beta}}_j}{r}}{\dfrac{\hat{\phi}}{n - \mathrm{tr}(\mathbb{F})}} \xrightarrow{D} \chi^2_r, \text{ where } n \to \infty.$$

And so, based on this distributional results we can perform Wald tests for significance of regression coefficients $\hat{\boldsymbol{\beta}}_j$ in the following form:

$$H_0 : \boldsymbol{\beta}_j = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_j \neq \mathbf{0}$$

Note: Tests for 2 nested models from which we want to choose the prefferable one are performed by ANOVA for GAM which represents extension of method introduced in section 2.7 and can be found in [8].

To summarize it, we introduce GAM as extension of GLM in which the linear predictors can also partly depend linearly on some unknown smooth functions. Regression coefficients are estimated by a penalized version of the method used to fit GLM, where an extra criterion has to be optimized to find the smoothing parameters.

5. The practical part

5.1. The problem formulation

In this chapter we demonstrate the practical usage of models which have been introduced in the previous chapters of this thesis. Generalized linear models as well as their extension GAM have various applications in all fields related with statistics. The non-life insurance is no exception where GLM is considered to be the best market practise in pricing and in reserving.[1]

We are going to demostrate the situation in which pricing and reserving are merged into the one. Such situation can occure for example, when we need compute expected losses connected with CASCO policy, the motor accidence insurence, in two years time horizont. However, this is also our task and to accomplish it we have to create models for:

- Demand Rate: Probability that particular policy will be renewed on the end of the first term.
- Cancel Rate: Probability of the middle term cancellation.
- Claims Frequency: Frequency of claim occurrence during one year period.
- Claim Severity: Severity of occurred claim.
- Days to Cancellation: Life expectence of policy in the case of middle term cancellation.

For these models we introduce following notation of their expected responses (the mean values of corresponding responses):

- Demand Rate: D,
- Cancel Rates for the first and the second term: C_1, C_2,
- Claims Frequencies for the first and the second term: F_1, F_2,
- Claims Severities for the first and the second term: S_1, S_2,
- Days to Cancellation for the first and the second term: R_1, R_2.

We assume that the Demand Rate, Cancel Rate, Claims Frequency, Claim Severity, Days to Cancellation are mutually independent random variables. Their distributions and specifics are presented in the corresponding sections devoted to the particular modelling stages.

Note: The terms in our case represent calendar years.

[1] In this thesis we are using terminology common in insurance practise.

On our desired result we look as on the expected written burning costs in two-year time horizon which can be expressed as:

$$BC = BC_1 + BC_2,$$

where BC_1 can be computed as:

$$BC_1 = C_1 \frac{R_1}{365} F_1 S_1 + (1 - C_1) F_1 S_1.$$

Note: $F_1 S_1$ represents netto premium of policy in the first year in case it hasn't been cancelled. Probabilty that policy is not cancelled during the first year is $(1 - C_1)$. Probability that policy has been cancelled during the first year is C_1 and then $\frac{R_1}{365}$ represents expected part of the year in which the policy has been valid.

And BC_2 can be computed as:

$$BC_2 = D.(1 - C_1).\left(C_2.\frac{R_2}{365}.F_2.S_2 + (1 - C_2).F_2.S_2\right).$$

Note: $D.(1 - C_1)$ is the probability that policy isn't cancelled during the first year and afterwards is renewed. Otherwise, it is analogical to the computation of BC_1.

The datasets on which the model are built comes from databases of Generali Insurance Group and are connected to the east european markets. Here, we present table of predictors used in final models.

Table 5.1 Predictors used in final stage of modelling

Categorical Predictors	*No. of Levels*	*Possible Values*
Policy.Anniversary.Month	12	1,2,…,12
Policy.Deductible.Group	4	1.No Deductible, 2.<=1%,2.<=2%,4.> 2%
Policy.New	2	Yes/No
Policy.Other.Drivers	2	Yes/No
Policy.Payment.Frequency	4	1,2,4,12
PolicyHolder.Bonus.Class	10	B0,B1,…,B7,M1,M2
PolicyHolder.Region	11	R01,R02,…,R11
Vehicle.Previous.owners	4	1,2,3,4
Continuous Predictors	***Units***	***Description***
Policy.Premium	Eur	the premium paid in current term
Policy.Previous.Premium	Eur	the premium paid in previous term
PolicyHolder.Age	Years	the age of policyholder on the beginning of term (company 0)
PolicyHolder.Latitude	Coordinates	the horizontal GPS coordinates
PolicyHolder.Longitude	Coordinates	the vertical GPS coordinates
PolicyHolder.Mileage.per.year	Miles	the average annual miles
Vehicle.Age	Years	the age of vehicle on the beginning of the term
Vehicle.Power	kW	the power of vehicle engine
Vehicle.Sum.Insured	Eur	the limit of policy coverage
Vehicle.Value.EUR	Eur	the vehicle price estimate in current term

Note: Complete list of available predictors with corresponding transformation for particular models can be found in attached ℝ code.

Such task enable us to introduce models with three types of response distribution, binomial, poisson and gamma (see section 2.1.) and serves as proper presentation of GLM/GAM capabilities. Resulting models can be afterwards used for computation of "optimal premium rates" by tools of optimalization.

For the technical part of this thesis we use the generalized additive modelling functions provided by statistical program ℝ, namely it's package *mgcv*. The main *gam* function from this library is very much like the often used *glm* function(following the method introduced in section 4.4.). The main difference is that the *gam* model formula can include smooth term, function *s*(), and there is number of options available for controlling automatic smoothness selection or for manual controlling model smoothness.

Individual models are created manually. We iteratively add particular predictors to the existing models. By method described in the following section, which is based on graphical illustration of model fit and effect of the new predictor in the single profile analysis, we determine way how to adjust existing model. For investigation of predictor importance and consequent reparametrization, we use graphical ilustration(as is the best market practise). Besides it's non-mathematical nature it is very usefull and brings the best results. For such graphical ilustrations we have created functions, which source codes can be found in attached ℝ-code.

Note: For reparametrization of the model also exist statistical algorithms, for example the stepwise algorithm, which can deliver statistically more precise models. However, for big datasets these algorithms usually propose models with big number of predictors and due to big correlation between predictors their modelled effects can be also illogical.

5.2. Demand model

In the last few years, due to econamical crisis, demand modelling became important practise in the insurance business. It is used as tool for expressing the price elasticity of policyholders and afterwards used for optimalization of prices on individual level. The response of such model has binomial type of distribution and as the link function is the most often used logit or probit. In our case, the model is determined by response variable:

$$Response.Demand := \begin{cases} 1 & , \text{ policy was renewed} \\ 0 & , \text{ policy was not renewed} \end{cases},$$

and as link function we use:

$$logit\,(\mu_i) := log\mu_i / (1 - \mu_i).$$

Note: The type of regression analysis is logistic regression.

Model construction is done iteratively and particular steps can be found in the attached ℝ code. However to ilustrate how the model was built, we demonstrate how the last predictor was added to the model.

The formula of current model is:

$$logit(Response.Demand) = log(Policy.Premium) + Vehicle.Power + s(PolicyHolder.Age,\ k = 7) + Vehicle.Age.T + Policy.Payment.Frequency + Deductible.Group + Policy.Anniversary.Month + PolicyHolder.Region. \quad (16)$$

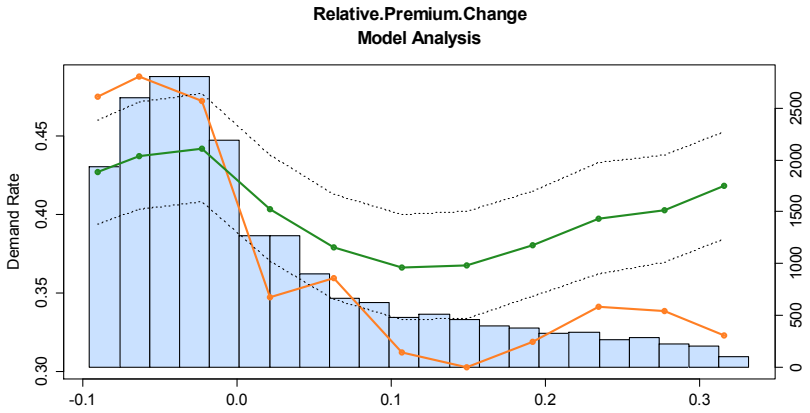
Note: Current model consists of predictors, listed above, described in Table 4.1. and *s(PolicyHolder.Age, k = 7)* represents thin plate spline with 7 knots for predictor *PolicyHolder.Age.*

We want to examine, if the extension of model by *Relative.Premium.Change* would bring statistically significant improvement of the current model. *Relative.Premium.Change* is predictor defined as:

$$Relative.Premium.Change := \frac{Premium - Previous.Premium}{Previous.Premium},$$

where *Previous.Premium* represents premium paid in the first term and *Premium* is the offered price, which is the policyholder supposed to pay in the second term.

Firstly, we are interested in how well the current model describes dependence between response and new predictor which we want to add. For this purpose we use model analysis of predictor *Relative.Premium.Change* for the current model, see Picture 1.:



Picture 1. Demand.Rate vs Relative.Premium Change for model DEM.10

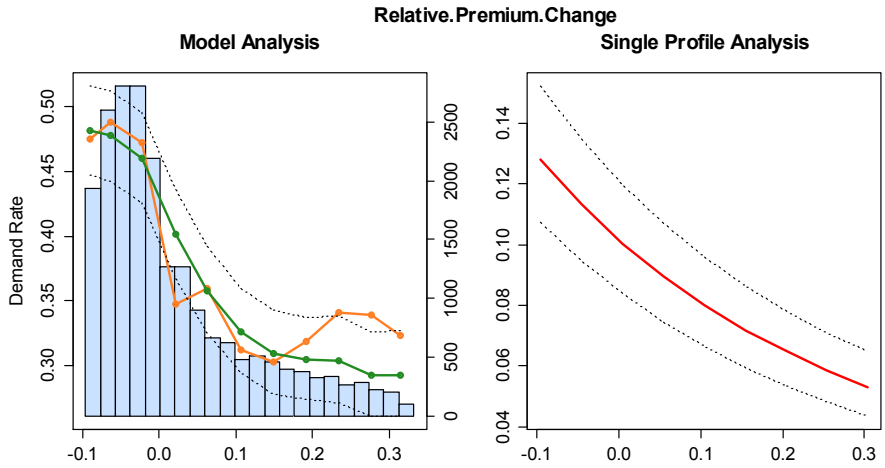
The green line represents how predictions of the existing model depend on the values of the new added predictor(relative premium change) and the orange line represents actual response. The primary *y*-coordinates represent means of these values and the secondary *y*-coordinates correspond to the counts for histogram of new predictor which values are the *x*-coordinates. Same notations will be also used in all next plots, if we didn't say otherwise.

Note: Functions created in ℝ enable to focus on specific values of studied predictors by setting manual axis scaling. We use manual scaling for the values of predictor among 2% and 98% percentile, other values have very low exposure.

From the plot in Picture 1. we can see that adding this new predictor will have very probably big impact on the model estimates. Also, predictor describing premium relative change is very important for determination of price elasticity of policyholders and so it's adding is very desirable. From bussiness logic, complicated spline for description of it's effect is not very proper and based on the shape of the dependence is not even invetibale. On the other hand, *log(Relative.Premium.Change+1)* may be proved to be proper transformation. Hence, as new model we propose:

$$\begin{aligned} logit(Response.Demand) &= log(Policy.Premium) \\ &+ Vehicle.Power + s(PolicyHolder.Age,\ k = 7) + Vehicle.Age.T \\ &+ Policy.Payment.Frequency + Deductible.Group + Policy.Anniversary.Month \\ &+ PolicyHolder.Region + log(Relative.Premium.Change + 1). \end{aligned} \tag{17}$$

To judge how big improvement was achieved, we take a look on Picture 2 produced by function *after.C,* which code can be found in attached ℝ code:



Picture 2. Demand.Rate vs Relative.Premium Change for last iteration of demand model

If we compare plots representing model analysis before (Picture 1.) and after (Picture 2.) adding the predictor, we can observe quite big improvement in the model ability to describe effect of relative premium change on the demand rate. We could argue, if the transformation was the best possible and based on the plot we could find functions to better describe effect of "bigger" relative premium changes. However, these "bigger" changes represents only very small proportion of the dataset and their exposure is very low and so corresponding estimates wouldn't be very reliable. Therefore, we should focus on "small" changes where, based on plots, is this transformation more than satisfing.

The second plot in Picture 2. serves for verification of the form of resulting transformation. It illustrates single profile analysis which represents how much would demand rate be varying for one particular policyholder, randomly chosen, in dependence on the values of relative premium change if all other characteristics stay constant.

However, graphical illustration can be somehow confusing and can be hardly seen as correct mathematical proof of suitability of model extension. So after each iteration we perform analysis of variance (ANOVA see section 2.7.) to compare new model with the old one.

Accordingly, in the current (last) iteration we set hypothesis:

$$H_0: \ logit(Response.Demand_i) = \mathbb{X}_i^*.\boldsymbol{\beta}^* \quad \& \quad H_1: \ logit(Response.Demand_i) = \mathbb{X}_i.\boldsymbol{\beta}, \quad \forall\, i \in \{1, \ldots n\},$$

where $\mathbb{X}_i^*$ represents matrix of predictors of model (16) (submodel) with corresponding regression coefficients $\boldsymbol{\beta}^*$ and $\mathbb{X}_i$ represents matrix of predictors of model (17) (full model) with corresponding regression coefficients $\boldsymbol{\beta}$

Note: We set the test to use χ^2 distribution since the scale parameter of binomial distribution is known.

The test produces following result:

Table 5.2 Anova for last iteration of demand rate model

model	*resid. e. df.*	*deviance*
(16)	25 413	28 359
(17)	25 412	26 962
difference:	0.99	1 397.30
p-value(χ^2):		***< 2.2e-16***

Therefore based on *p-value* (<2.2e-16) we can consider the extended model for more prefferable.

Subsequently, we perform *Wald test* (see section 4.4.) to check if all used predictors are after adding new predictor still statistically significant.

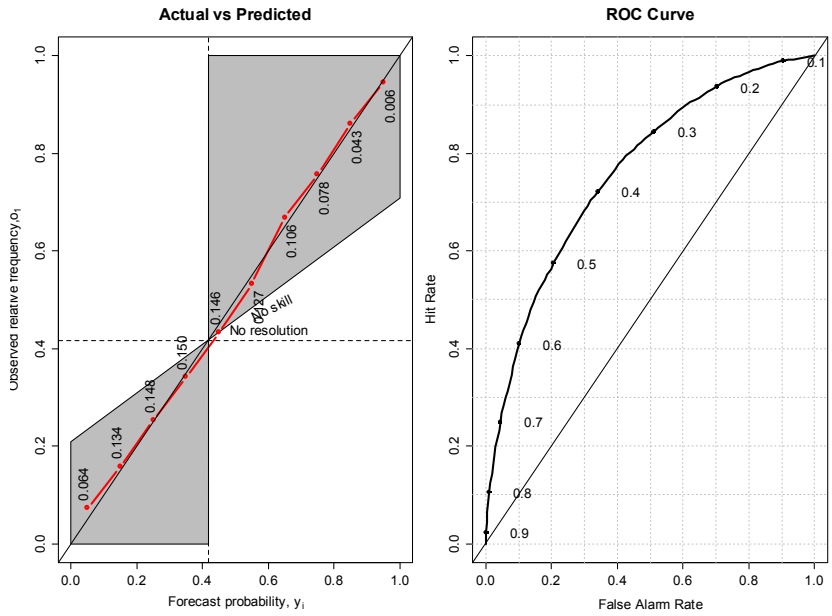
Table 5.3 *Wald test* for final demand rate model

Predictors	*df*	χ^2	*p-value*
log(Policy.Premium)	1	53.62	2.44e-13
Vehicle.Power	1	204.61	<2e-16
Vehicle.Age.T	1	176.8	<2e-16
Policy.Payment.Frequency	3	527.31	<2e-16
Policy.Deductible.Group	3	604.6	<2e-16
Policy.Anniversary.Month	11	483.64	<2e-16
PolicyHolder.Region	9	565.75	<2e-16
log(Relative.Premium.Change + 1)	1	274.86	<2e-16
Approximate significance			
Smooth terms	***e. df***	***ref. df***	χ^2	***p-value***
s(PolicyHolder.Age)	*5.899*	5.994	226.1	<2e-16

In the presented table *df* represents the degrees of freedom of corresponding χ^2 statistic for particular predictor (analogically *e.df* represents effective degrees of freedom).

Based on the particular *p-values* we can consider all used predictors for significant. Same notation of this parameters will be also used in all next tables describing used predictors of final models, if we didn't say otherwise.

Let's assume that this model is considered to be sufficient and we want to verify, if it's quality is adequate for practical usage. For this purposse we analyze results of final fit by graphical illustration:



Picture 3. Fit Analysis for Demand.Rate

Actual vs. Predicted plot provides a visual comparison between the actual value in the dataset and the corresponding estimated response and shows how well the model fits the data. The *x*-coordinates are as before means of estimated responses and the *y*-coordinates are means of corresponding actual values. A diagonal line represents the points where predicted and actual values are the same. For a perfect fit, all the points would be on this diagonal. Uncertainty is described as the vertical distance between this point and the diagonal line. This plot suggests that our model fits data in very good way.

The *ROC* graph illustrates relative trade-offs between benefits(hits) and costs(false alarms). Thus, the *ROC* curve plots the false alarm rate against the hit rate for a probablistic estimates for a range of thresholds(10-quantiles). In these tresholds we set cut-off probabilities when higher values of $Response.Demand_i$ are undestood as positive responses, i.e. policies are renewed. For the specific cut-off probabilities we determine:

$$Hit\ Rate = \frac{Positives\ correctly\ classified}{Total\ positives} = \frac{TP}{TP + FN},$$

$$False\ Alarm\ Rate = \frac{Negatives\ incorrectly\ classified}{Total\ negatives} = \frac{FP}{FP + TN}.$$

Notice, there are four possible outcomes for binomial response. If the predicted response is that the policy is renewed and the policy has been actually renewed then it is called a true positive (*TP*). However, if the policy has been actually cancelled on anniversary then it is said to be a false positive (*FP*). Conversely, a true negative (*TN*) has occurred when both the predicted response and the actual response suggest that the policy is not renewed, and false negative (*FN*) is when the predicted response is that the policy is not renewed while actually renewed is.

We use same notation also in the following plots analyzing the fit results, if didn't say otherwise.

The area under the *ROC* curve *AUC* is understood as a measure of a estimate's accuracy. A measure of 1 would indicate a perfect model. A measure of 0.5 would indicate a random forecast. The *AUC* is related to the (in practice frequently used) *Gini* coefficient G_1 by the formula

$$G_1 = 2\,AUC - 1.$$

In our model area under the *ROC* curve is 0.76. As a rough guide for classifying the reliability of a model based on *AUC* is in the practise often used scale:

- 0.90-1.00 = excellent,
- 0.80-0.89 = good,
- 0.70-0.79 = fair,
- 0.60-0.69 = poor,
- 0.50-0.59 = fail.

Therefore, our model would be considered to be "fair", at separating case where the policy is renewed from case where don't. Estimates from our model hence can be considered to be reliable enough, what together with the fact that effects of particular response makes businness sense leads to conclussion that created demand model is good enough for usage in practice.

Single profile analysis for all used predictor, which can give very good visual idea of particular effects, can be demonstrated by using attached ℝ code. Now we interprate model results introduced in table 5.4 with respect to the business logic:

- *Policy.Payment.Frequency*- with increasing frequency of payments renewal rate is also increasing probably due to lower immediate payment in the moment of renewal.
- *Policy.Deductible.Group*- with increasing level of deductibles renewal rate is also increasing probably due to lower price level of such policies.
- *Policy.Anniversary.Month*- during winter is renewal rate lower in the opposite to the summer when it is higher what corresponds with market experience.
- *PolicyHolder.Region*- big differences in the renewal rates corresponds with standards of living in the particular regions. In richer regions we expect higher renewal rate.
- *Vehicle.Age.T*- for older cars is renewal rate higher due to worse price options on the market for old cars.
- *Relative.Premium.Change*- with increasing price ranewal rate is decreasing.

- *Policy.Premium*- with increasing level of offered premium renewal rate is decreasing.
- *Vehicle.Power*- with increasing power of vehicle premium renewal rate is decreasing.
- *PolicyHolder.Age*- is used spline as is market practise due to big changes in renewal rates for different ages

Table 5.4 Summarizaton of predictors of final demand rate model

Categorical Predictors	***Levels***	***Regression Coefficients***	
		Estimates	***Std. Errors***
INTERCEPT		*-0.7601354*	0.2615817
Policy.Payment.Frequency	*2*	*0.845097*	0.0538545
	4	*0.7980229*	0.0365623
	12	*0.3931987*	0.0719917
Policy.Deductible.Group	*2.<=1%*	*0.5791014*	0.033996
	3.<=2%	*1.1341826*	0.0529868
	4.> 2%	*1.3287439*	0.1781191
Policy.Anniversary.Month	*2*	-0.1325529	0.0883699
	3	0.0196991	0.0850213
	4	0.2723835	0.0798519
	5	0.5432671	0.0749716
	6	0.6487246	0.0751698
	7	0.7388759	0.0723102
	8	0.787978	0.074717
	9	0.5869404	0.0787194
	10	0.431842	0.0763627
	11	0.0772644	0.0762852
	12	-0.2360626	0.0780358
Region	*R02*	1.6042869	0.2186268
	R03	1.2051194	0.0816404
	R04	0.1897306	0.1501979
	R05	1.6822947	0.1019669
	R06	0.6194212	0.1400566
	R07	1.7370018	0.0918404
	R08	1.2447694	0.0868298
	R10	*1.5204394*	0.0974941
	R11	*0.9830114*	0.0852183
Vehicle.Age.T	*Older Car*	*0.9583068*	0.072071
Continuous predictors	***Transformations***	***Estimates***	***Std. Errors***
Relative.Premium.Change	*log(x+1)*	*-2.6348671*	0.1589304
Policy.Premium	*log(x)*	*-0.3287215*	0.0448928
Vehicle.Power	*identity*	*-0.0138619*	0.0009691
PolicyHolder.Age	*thin plate spline with 7 knots*		

Note: The list of available predictors together with plots ilustrating impact of the particular predictor on the response for this model can be found in R code.

5.3. Mid term cancellation models

Although mid term cancellation models usually play suporting role in the process of pricing, they are still very important for calculations of the life time values of policies. And while the demand rates can be increased or decreased by pricing strategy, mid term cancellations ussualy depend on factors which can be hardly influenced. However, we can find segments, like segment of old cars, for which the probability of cancellation is much higher. This recognitions can be crucial for pricing strategies, because in such segments we should not rely on very insecure future incomes from these policies, but ratherly set prices in such a way that they will bring profit in shorter time period.

The probability of mid term cancellation can be described by model for cancel rate with response defined as:

$$Response.Cancellation := \begin{cases} 1 & , \text{ policy was cancelled} \\ 0 & , \text{ policy was not cancelled} \end{cases}.$$

Type of response distribution is binomial and as link function we are going to use:

$$logit := log\mu_i / (1 - \mu_i).$$

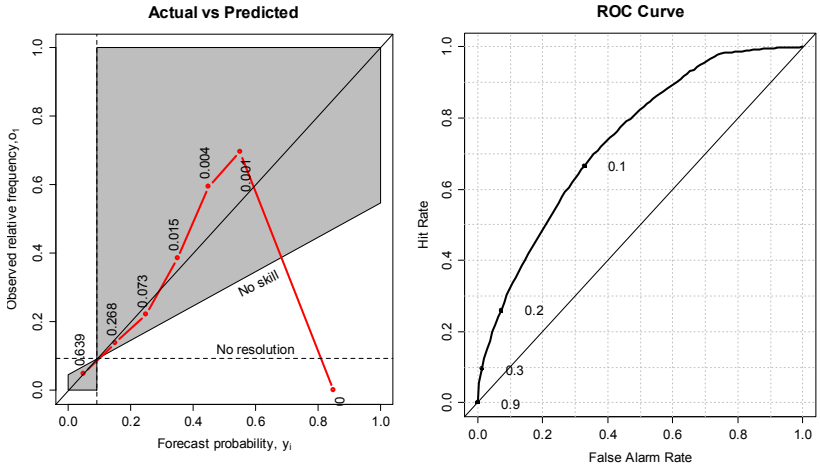
Subsequently, we perform *Wald test*.

Table 5.5 *Wald test* for final cancel rate model

Predictors	***df***	χ^2	***p-value***
log(Policy.Premium + 1)	1	175	<2e-16
Vehicle.Age	1	318.5	<2e-16
Policy.Payment.Frequency	3	2151.6	<2e-16
Policy.New	1	1135.2	<2e-16
Approximate significance			
Smooth terms	***e. df***	***ref. df***	χ^2 ***p-value***
s(PolicyHolder.Age)	*3.451*	3.79	234.7 <2e-16

Note: Based on the particular *p-values* we can consider all predictors for significant.

Analysis of fit results for Cancel Rate model



Picture 4. Fit Analysis for Cancel.Rate

The first plot suggests that our model fits data in good way in segmets with satisfying exposure. Resolution of the model is quite low due to small average cancel rate in the observed dataset. However the area under the *ROC* curve is 0.74, therefore our model can be considered to be "fair" at separating cases, when policy is cancelled from cases when don't. And so as well as for demand model holds, that estimates can be considered to be reliable enough.

Table 5.6 Summarizaton of predictors of final cancel rate model

Categorical Predictors	*Levels*	*Regression Coefficients*	
		Estimates	*Std. Errors*
INTERCEPT		*-6.372869*	0.148945
Policy.Payment.Frequency	*2*	*2.393604*	0.083871
	4	*2.857504*	0.077186
	12	*3.675746*	0.083458
Policy.New	*Yes*	*-0.810422*	0.024053
Continuous predictors	***Transformations***	***Estimates***	***Std. Errors***
Policy.Premium	*log(x+1)*	*0.265074*	0.020036
Vehicle.Age	*identity*	*0.100235*	0.005616
PolicyHolder.Age	*thin plate spline with 5 knots*		

Interpretation of model results:

- *Policy.Payment.Frequency*- with increasing frequency of payments cancel rate is also increasing probably due to nonpayment.
- *Policy.New*- policyholders, who stay with us one term is more probable to stay another term.
- *Policy.Premium*- with increasing level of premium cancel rate is increasing.
- *Vehicle.Age*- with increasing vehicle age cancel rate is also increasing probably due to selling the vehicle.
- *PolicyHolder.Age*- is used spline as is market practise due to big changes in renewal rates for different ages

In case the policy would be cancelled we need model enabling us to express part of the year during which we would be exposed to the risk. Response for such model can be defined as count of corresponding days as

$$Response.Days.to.Cancellation := k \quad , k \in \{1, \ ..., 364\}.$$

We assume that type of response distribution is Poisson and as the link function we use:

$$log := log\mu_i.$$

<u>*Note*</u>: Such model can be also used for modelling refund proportion representing part of the premium returned to the client after cancelling the policy.

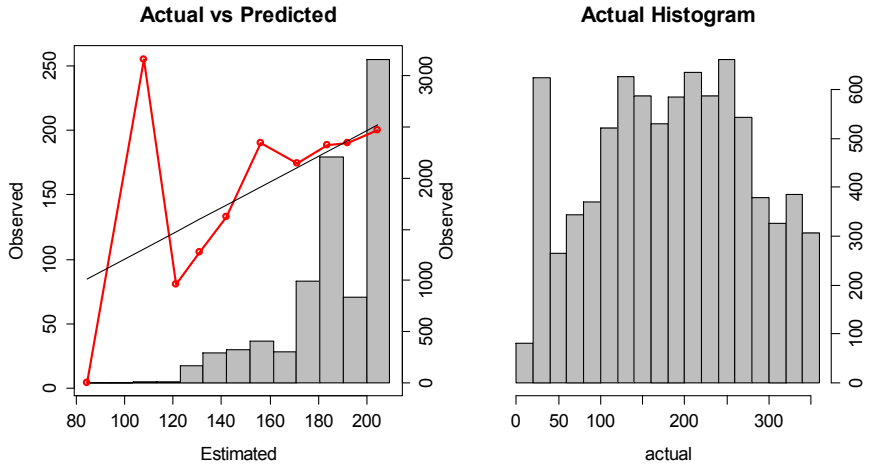
Table 5.7 *Wald test* for final days to cancellation model

Predictors	*df*	χ^2	*p-value*
Policy.Premium	1	1206	<2e-16
Policy.Deductible.Group	3	1389	<2e-16
Policy.Payment.Frequency	1	10361	<2e-16
Policy.New	1	2802	<2e-16

Note: Based on the particular *p-values* we can consider all predictors for significant. We haven't used any splines and so propossed model belongs to the *GLM*.

Analysis of fit results for Day to Cancellation model

Type of distribution of response for this model is not binomial and so we can not use *ROC*, but we are still able to create Actual vs. Predicted plot. However it has slightly different formatting. Also, we have added histogram of actual responses.



Picture 5. Fit Analysis for Days.to.Cancellation

From these plots we can see, that our model isn't very reliable. On the one hand in the Actual vs. Predicted it is doing well in segments with high exposure, but histograms show that distributions of actual and estimated differ significantly. Actual histogram suggests that the recognition capacities of our model are not sufficient. However this model can be considered to be adequate for our purposses due to low cancel rate, which diminishes it's importance.

Table 5.8 Summarizaton of predictors of final days to cancellation model

Categorical Predictors	*Levels*	*Regression Coefficients*	
		Estimates	*Std. Errors*
INTERCEPT		*5.18*	*0.00224*
Policy.Payment.Frequency.T	*12*	*-0.2699*	*0.002651*
Policy.Deductible.Group	*2.<=1%*	*0.07373*	*0.002103*
	3.<=2%	*0.07615*	*0.002549*
	4.> 2%	*0.02261*	*0.005947*
Policy.New	*Yes*	*0.08912*	*0.001684*
Continuous predictors	***Transformations***	***Estimates***	***Std. Errors***
Policy.Premium	*identity*	*-6.39E-05*	0.00000184

Interpretation of model results:

- *Policy.Payment.Frequency.T*- policyholders who are paying monthly are more probable to cancel the policy sooner.
- *Policy.New*- policyholders who stayed with us one term are more probable to cancel policy later in the next one.
- *Policy.Premium*- with decreasing level of premium policies are more probable to cancel later.
- *Policy.Deductible.Group*- with increasing level of deductibles policies are more probable to cancel later.

Note: The list of available predictors together with plots ilustrating impact of the particular predictor on the response for both models can be found in ℝ code.

5.4. Risk models

Price for received premium is transfer of specific risks to an insurance company. Such risks can be evaluated by the expected loss. In order to express this loss we compute aggregate costs, repressenting the sum of all claims in the period of time during which the policy covered given risks. The most datasets, as well as our, include number of claims on policy level together with the corresponding time during which policy was exposed to the risk. Since we are interested in number of claims which can occur in one year time period, we define:

$$Claims.Frequency := \frac{Claims.Number}{Policy.Exposure}.$$

Note: Usually datasets are created based on informations collected to some moment in time. However, we have to take to consideration also time delay with which are claims reported to the insurer. Hence we have to consider extra loading(some coefficients) for unreported claims. It is common practise to increase the number of known claims with respect to the time period in which claim incuared. This loading is known as *IBNR*(Incuared but not reported). Therefore, in this chapter we use variables *Claims.Number.IBNR* and *Claims.Severity.IBNR* which content multiplicative coefficients increasing the number and severity of reported claims by multiplying them with corresponding loadings.

Response of model for claims frequency is defined as:

$$Claims.Frequency := \frac{Claims.Number \ . \ Claims.Number.IBNR}{Policy.Exposure}.$$

Althogh, such defined response is not an integer we still assume that type of it's distribution is Poisson and as link function we use

$$log := log\mu_i.$$

Note: Recall, the estimation process of GLM, as well as process of GAM, simply mean that we want to solve equations (8) based on relation between regression coefficients

and cannonical parameter. And it is possible to solve them, without any condition such that values of observations should be integers.

Table 5.9 *Wald test* for frequency model

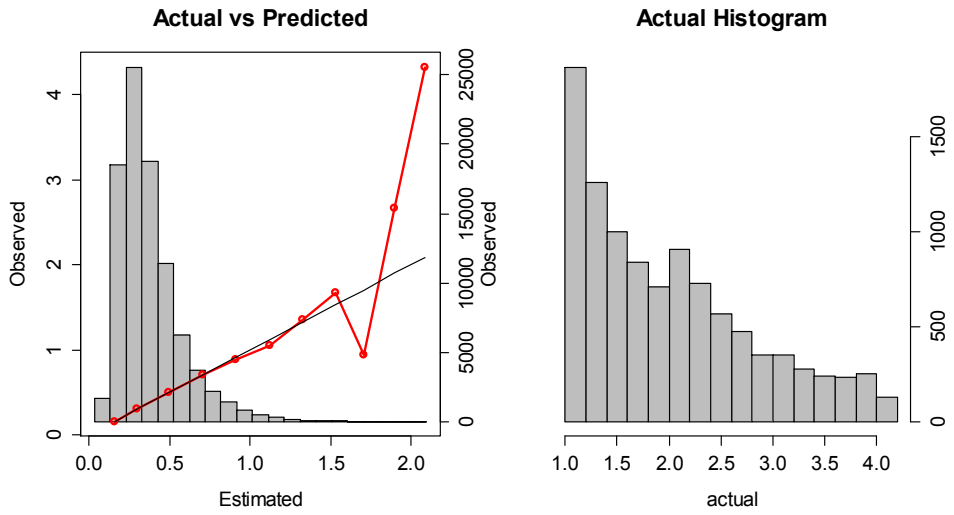
Predictors	*df*	χ^2	*p-value*
log(Vehicle.Value.EUR + 1)	1	413.38	<2e-16
PolicyHolder.Mileage.per.year.T	1	30.51	3.33e-08
Policy.Deductible.Group	3	1542.5	<2e-16
PolicyHolder.Bonus.Class.T	6	586.56	<2e-16
Policy.Payment.Frequency	3	58.08	<1.51e-12
Vehicle.Previous.owners.T	1	46.57	<8.84e-12
Policy.Other.Drivers	1	28.08	<1.16e-07

Approximate significance

Smooth terms	*e. df*	*ref. df*	χ^2	*p-value*
s(PolicyHolder.Age)	*4.372*	4.966	253.2	<2e-16
s(PolicyHolder.Latitude, PolicyHolder.Longitude)	*31.58*	38.979	418	<2e-16

Note: Based on the particular *p-values* we can consider all predictors for significant.

Analysis of fit results for Frequency model



Picture 6. Fit Analysis for Frequency model

The Actual vs. Predicted plot suggests that our model fits data in very good way in all segmets with almost any exposure and has big recognition capabilities, when it enables to identify policyholders with low claims frequency as well as policyholders with very high claims frequency. We also plotted Actual histogram from which we excluded the nonclaimers to check if the shapes of distributions correspond. The distribution do not indicates any serious problems, e.g. heavy tails. And so model can be considered to be reliable.

Table 5.10 Summarizaton of predictors of final frequency model

Categorical Predictors	*Levels*	*Regression Coefficients*	
		Estimates	*Std. Errors*
INTERCEPT		*-4.23035*	*0.18341*
Policy.Payment.Frequency	*2*	*0.05357*	*0.02873*
	4	*0.12856*	*0.01968*
	12	*0.22237*	*0.03985*
Policy.Deductible.Group	*2.<=1%*	*-0.62584*	*0.01878*
	3.<=2%	*-0.83871*	*0.02747*
	4.> 2%	*-1.34355*	*0.09948*
PolicyHolder.Bonus.Class.T	*B1-B3*	*-0.17174*	*0.01895*
	B4	*-0.32795*	*0.02475*
	B5	*-0.5164*	*0.03067*
	B6	*-0.57504*	*0.04269*
	B7	*-0.86317*	*0.1167*
	M1-M2	*0.18136*	*0.04003*
Policy.Other.Drivers	*Yes*	*0.13133*	*0.02478*
Vehicle.Previous.owners.T	*2+*	*0.15232*	*0.02232*
PolicyHolder.Mileage.per.year.T	*10000+*	*-0.13536*	*0.02451*
Continuous predictors	***Transformations***	***Estimates***	***Std. Errors***
Vehicle.Value.Eur	*log(x+1)*	*0.38662*	0.01902
PolicyHolder.Age	*thin plate spline with 7 knots*		
PolicyHolder.Latitude, PolicyHolder.Longitude	*two dimensional thin plate spline with 50 knots*		

Interpretation of model results:

- *Policy.Payment.Frequency*- with increasing frequency of payments claims frequency is also increasing what corresponds with market experience.
- *Policy.Deductible.Group*- with increasing level of deductibles claims frequency is decreasing due to no participation of insurer on small claims.
- *PolicyHolder.Bonus.Class.T*- with worsening driving history claims frequency is increasing.
- *Policy.Other.Drivers*- policies which cover also other drivers have higher claims frequency.
- *PolicyHolder.Mileage.per.year*-policyholders with more driven miles have lower claims frequency.
- *Vehicle.Value.Eur*- with increasing vehicle value claims frequency is also increasing.
- *PolicyHolder.Age*- is used spline as is market practise due to big changes in renewal rates for different ages
- *PolicyHolder.Latitude,PolicyHolder.Latitude*- is used two-dimensional spline to create risk(heat) map.

Severity of claims incuaring with introduced frequencies can be defined as average amount of incuared claims and so taking into account IBNR loadings can be written as:

$$Claims.Severity := \frac{Claim.Severity \ . \ Claim.Severity.IBNR}{Claims.Number \ . \ Claims.Number.IBNR}.$$

Type of it's distribution, based on the corresponding histogram of actual response, Picture 7, can be identified as gamma distribution. As link function we are also going to use

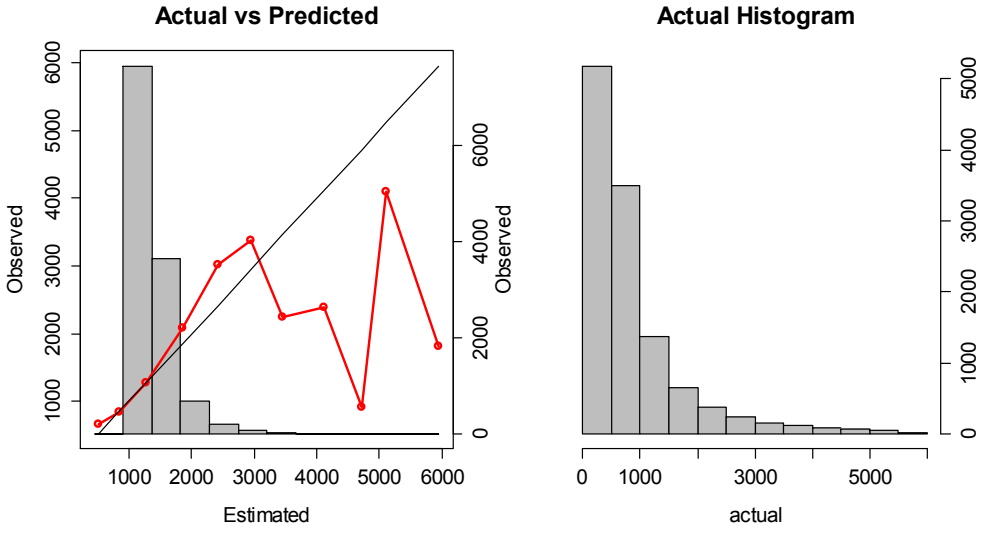
$$\log := \log \mu_i.$$

Table 5.11 *Wald test* for severity model

Predictors	*df*	*F*	*p-value*
log(Vehicle.Sum.Insured + 1)	1	262.59	<2e-16
PolicyHolder.Region	10	7.181	2.15e-11
Vehicle.Power.T	1	31.002	2.63e-08
Vehicle.Power	1	15.095	1.03e-05
Vehicle.Age	1	133.847	<2e-16
Policy.New	1	85.908	<2e-16

Note: Based on the particular *p-values* we can consider all predictors for significant. We haven't used any splines and so propossed model belongs to GLM.

Analysis of fit results for Severity model



Picture 7. Fit Analysis for Severity model

The Actual vs. Predicted plot suggests that our model fits data in good way in segmets with satisfying exposure . We also plotted Actual histogram to check if the shapes of distributions correspond and if we do not neglect heavy tails for claims severities by choice of gamma distribution. The distributions do not indicates any serious problems.

And so model can be considered to be reliable.

Note: We haven't used any splines and so propossed model belongs to the GLM. We did it, because of possible overfitting caused by small size of dataset and possible big randomness in response variable due to fact that severity of claims is variable varying in quite big way from case to case and so can not be "perfectly" estimated. However, during the process of fitting we have encountered the situation, when we needed to use some form of automatic parametrization of predictors and so it was preferable to use methods of segmented generalized linear models. Concretely, we want include *Vehicle.Power* as predictor to the model because models for severity usually include this variable. However, adding this predictor didn't have expected effect due to fact that severities of claims increased rapidly from some level of this predictor, but to this point it didn't have such effect. Such point represents breakpoint and so we use algorithm desribed in chapter 3. For it's localization the propossed value was 75, and so the new predictor

$$Vehicle.Power.T = (Vehicle.Power - 75)_+$$

was added to the model and both *Vehicle.Power* and *Vehicle.Power.T* became statistically significant.

Table 5.12 Summarizaton of predictors of final severity model

Categorical Predictors	*Levels*	*Regression Coefficients*	
		Estimates	*Std. Errors*
INTERCEPT		*2.16266*	0.297507
Region	*R02*	-0.164668	0.125111
	R03	-0.191607	0.058521
	R04	-0.251708	0.107425
	R05	-0.336999	0.065714
	R06	-0.081864	0.091791
	R07	-0.344235	0.061558
	R08	-0.180122	0.061199
	R09	-0.150787	0.097942
	R10	*-0.143919*	0.065783
	R11	*-0.204905*	0.061755
Policy.New	*Yes*	*-0.174832*	0.018863
Continuous predictors	***Transformations***	***Estimates***	***Std. Errors***
Vehicle.Sum.Insured	*log(x+1)*	*0.552542*	0.034098
Vehicle.Power	*identity*	*-0.004196*	0.00108
Vehicle.Power	*max(x-75,0)*	*0.006394*	0.001148
Vehicle.Age	*identity*	*0.07509*	0.00649

Interpretation of model results:

- *PolicyHolder.Region*- regional differences are to be expected.
- *Policy.New*- policyholders in the first year cause claims with smaller severity than in the next year.

- *Policy.Deductible.Group*- with increasing sum insured claims severity is also increasing due to higher limit of coverage.
- *Vehicle.Power*- vehicles with small or big engine power causes claims with higher severity than vehicles with an average power.
- *Vehicle.Age*- with increasing vehicle age severity is also increasing.

Expected Burning costs, aggregate expected loss during one year, for our risk models on the policy level are then given by:

$$E[Aggreage\ loss] = E[Claims.Frequency].E[Claims.Severity].$$

5.5. Results

We have prepared all models in order to be able to complete the task which we have stated in the section 6.1. and that compute expected losses connected with CASCO policy in two years time horizont. To demostrate possible results and to recapitulate used predictors we introduce example computed for random policyholder profile. Attached ℝ code can provide results for any profile selected by user.

Table 5.13 Values of predictors for random profile

Predictors	*Random Profile*	
	First Year	*Second Year*
Policy.Anniversary.Month	5	5
Policy.Deductible.Group	2.<=1%	2.<=1%
Policy.New	Yes	No
Policy.Other.Drivers	No	No
Policy.Payment.Frequency	4	4
Policy.Premium	340	350
Policy.Previous.Premium		340
PolicyHolder.Age	35	36
PolicyHolder.Bonus.Class	B5	B6
PolicyHolder.Latitude	45	45
PolicyHolder.Longitude	26	26
PolicyHolder.Mileage.per.year	11000	11000
PolicyHolder.Region	R01	R01
Vehicle.Age	1	2
Vehicle.Power	80	80
Vehicle.Previous.owners	1	1
Vehicle.Sum.Insured	8000	8000
Vehicle.Value.EUR	10000	9000

Note: The gray color of fill sugests that the value of predictor can change between the first and the second term, other are considered to be constant.

Model Results for such values of predictors:

Table 5.14 Model results for random profile

1 Year	*Cancel Rate*	*6.63%*
	Days to Cancellation	*204.97*
	Claims Frequency	*18.78%*
	Claim Severity	*833.04*
	Demand Rate	*23.81%*
2 Year	*Cancel Rate*	*14.90%*
	Days to Cancellation	*187.59*
	Claims Frequency	*16.73%*
	Claim Severity	*1069.56*

Written burning costs, as were stated in the section 6.1., for the first year and the second year:

$$BC_1 = C_1.\frac{R_1}{365}.F_1.S_1 + (1 - C_1).F_1.S_1$$

$$BC_2 = D_2.(1 - C_1).\left(C_2.\frac{R_2}{365}.F_2.S_2 + (1 - C_2).F_2.S_2\right)$$

Evaluation for demonstrative model results gives:

$$BC_1 = 151.94 \text{ Eur}$$

$$BC_2 = 36.89 \text{ Eur}$$

Note: Burning costs for the second year are diminished by probability, that policy is even "alive" in the second term, otherwise our expected costs for the second term would be:

$$\left(C_2.\frac{R_2}{365}.F_2.S_2 + (1 - C_2).F_2.S_2\right) = 165.95 \text{ Eur}$$

To sum it up, we propose extension of GLM approach, typically used in insurance, to investigate the risks connected with non-life policy. On the one hand, we are able to achieve more precise model fits by introducing segmented generalized linear models and GAM. On the other hand, we also propose method for calculation of burning costs, which takes to consideration not only risk factors for one year time period but it focus on probable written costs during the expected life of the policy in two years' time horizon. This horizon could be theoretically extended for the infinite period but due to volatility of the market and the small portion of non-life CASCO policies surviving more than two years we choose this variant. However, in more stable environment and for other products this method can be analogically extended. We believe that potential of this method lies in creating new pricing strategies using the processes of optimization, when we could optimize the expected written margin as the function of offered premium in the first term with respect to the expected life of the policy.

Bibliography

[1] Andreas Brezger, Stefan Lang. 2012. Bayesian P-Splines. Journal of Computational and Graphical Statistics.

[2] Helmut Küchenhoff. 1996. An exact algorithm for estimating breakpoints on segmented generalized linear models. Paper 27.

[3] Jeffrey S. Racine. 2013. Primer on Regression Splines.

[4] John Nelder, Robert Wedderburn. 1972. Generalized linear models. Journal of the Royal Statistical Society, Series A (Royal Statistical Society) 135 (3): 370–384

[5] Memmedagha Memmedov, Rabia Ece Omay. 2013. Regression Models with Thin Plate Spline.

[6] MSc. Econ. 1995. Maximum-Likelihood Estimation, Mathematical Statistics.

[7] Simon N. Wood. 2006. Generalized Additive Models: An Introduction with R. CRC/ Chapmand and Hall.

[8] Trevor Hastie, Robert Tibshirani. 1999. Generalized Additive Models. Chapmand and Hall.

[9] Vito Muggeo. 2003. Estimating regression models with unknown break points. Statistics in Medicine. p 3055–3071.

Appendix

I. Consistency of maximum likelihood estimator

In order to derive an maximum likelihood estimator we would expect, that we have to make an assumption about the type of distribution from which response originates. However, the distributions can often differ without affecting the form of the maximumum likelihood estimator. Hence, the general theory of maximum likelihood estimation can be developed without reference to a specific distribution of the response variable. In order to reveal important characteristics of the likelihood estimators, we firstly investigate the properties of the corresponding log-likelihood function.

For simplicity of presentation we consider the case where θ is the sole parameter common for $Y_1, \ldots, Y_n$ of and so for log-likelihood function we have $l_{\mathbf{y}}(\boldsymbol{\theta}, \phi) := l_{\mathbf{y}}(\theta, \phi)$, where $\mathbf{y}$ is as ussual vector of observations of random variable $\mathbf{Y} = (Y_1, \ldots, Y_n)$. In seeking for the estimate of the parameter θ, we regard to it as to an argument of the log-likelihood function while the response is considered to have fixed values. However, in analysing the statistical properties of log-likelihood function, we return the role of random compoment to the response (we use insted of vector of fixed values $\mathbf{y}$ random sample $\mathbf{Y}$). This "randomness" is consequently transferred to the cannonical parameter maximising likelihood. Hence, $\hat{\theta}$ becomes estimator with statistical properties.

Remind

$$l_{\mathbf{Y}}(\theta, \phi) = \sum_{i=1}^{n} l_{Y_i}(\theta, \phi),$$

from which by multiplying both sides of equation by $\frac{1}{n}$ we get

$$\frac{1}{n}.l_{\mathbf{Y}}(\theta, \phi) = \frac{1}{n}.\sum_{i=1}^{n} l_{Y_i}(\theta, \phi).$$

For any value of θ, this represents a sum of mutually independent random variables with not only same type of distribution, but also with the exact same distribution, because of distribution of $Y_1, \ldots, Y_n$ are determined by same cannonical parameter. Hence, the Law of large numbers can be applied to the previous equation

$$\frac{1}{n}.l_{\mathbf{Y}}(\theta, \phi) \xrightarrow{P} E\, l_{Y_i}(\theta, \phi)\,, \text{ where } n \to \infty.$$

As we have shown in section 2.3., the expression

$$E\,\frac{1}{n}.\sum_{i=1}^{n} l_{Y_i}(\theta, \phi)$$

is maximal in the true value of cannonical parameter θ, for now denotate it as θ_0.

Therefore

$$\frac{1}{n}.l_{\mathrm{Y}}(\theta, \phi) \xrightarrow{P} E\,\frac{1}{n}\sum_{i=1}^{n} l_{Y_i}(\theta, \phi), \text{ where } n \to \infty$$

implies, that the estimated value $\hat{\theta}$, which maximize $\frac{1}{n}.l_{\mathrm{Y}}(\theta, \phi)$ converges under some assumptions of regularity to the true value of cannonical parameter θ_0. This leads to the consistency of maximum likelihood estimation. And so, in the process of estimation the distribution of $\hat{\theta}$ becomes increasingly concentrated around the true value of cannonical parameter θ_0. Maximum likelihood estimators are often not unbiased, but their consistency implies asymptotic unbiasedness, as dataset size tends to infinity.

Note: To show that $\hat{\theta}$ converges to the true value of cannonical parameter θ_0 in some well ordered manner as $n \to \infty$ requires an assumption of some regularity. For example, we need to be able to assume, at least, that if θ_1 and θ_2 are "close", then $l_{\mathrm{y}}(\theta_1, \phi)$ and $l_{\mathrm{y}}(\theta_2, \phi)$ are also "close". Luckily, in vast majority of practical situations such conditions hold.

A fundamental result is, that as the sample size increases, the likelihood function divided by the sample size tends to stabilise in such sense that it converges in probability at every point in its domain to a constant function. This leads to the recognition, that estimates by maximum likelihood functions are for big datasets very stable. On the other hand this method achieves poor results for small datasets.

For simplicity of explanation, the above argument dealt only with a single parameter θ and responses were independent observations of a random variables from one distribution. In fact, consistency holds in much more general circumstance: for vector parameters and non-independent data which do not necessarily all come from the same distribution.

II. Large sample distributions of maximum likelihood estimators

To obtain the large sample distribution of the maximum likelihood estimator $\hat{\theta}$, we express a Taylor expansion of the log likelihood function around the true value of cannonical parameter θ, θ_0:

$$l_{\mathrm{Y}}(\theta) \tilde{=} l_{\mathrm{Y}}(\theta_0) + \frac{\partial l_{\mathrm{Y}}(\theta_0)}{\partial \theta}.(\theta - \theta_0) + \frac{1}{2}.\frac{\partial^2 l_{\mathrm{Y}}(\theta_0)}{\partial \theta^2}.(\theta - \theta_0)^2 + \frac{1}{6}.\frac{\partial^3 l_{\mathrm{Y}}(\theta_0)}{\partial \theta^3}.(\theta - \theta_0)^3 + \cdots.$$

In pursuing the asymptotic distribution of the maximum likelihood estimator we can concentrate upon a quadratic approximation which is based on the first three terms of this expansion. The reason is, as we have mentioned above, that the distribution of the estimator becomes increasingly concentrated in the close distance of the true value of cannonical parameter θ_0 as the size of the sample increases. Therefore, the quadratic

approximation becomes increasingly accurate for big datasets in which we are interested.

The quadratic approximation at the point θ_0 is

$$l_Y(\theta) = l_Y(\theta_0) + \frac{\partial l_Y(\theta_0)}{\partial \theta}.(\theta - \theta_0) + \frac{1}{2}.\frac{\partial^2 l_Y(\theta_0)}{\partial \theta^2}.(\theta - \theta_0)^2.$$

It's derivative with respect to θ is

$$\frac{\partial l_Y(\theta)}{\partial \theta} = \frac{\partial l_Y(\theta_0)}{\partial \theta} + \frac{\partial^2 l_Y(\theta_0)}{\partial \theta^2}.(\theta - \theta_0).$$

Using the fact, that $\frac{\partial l_Y(\theta)}{\partial \theta}$ evalueted in maximum likelihood estimator $\hat{\theta}$ must be eqaual to 0, we find that evalatution of expression in $\hat{\theta}$ leads to

$$0 = \frac{\partial l_Y(\theta_0)}{\partial \theta} + \frac{\partial^2 l_Y(\theta_0)}{\partial \theta^2}.\left(\hat{\theta} - \theta_0\right).$$

It can be rewritten as:

$$\sqrt{n}.\left(\hat{\theta} - \theta_0\right) = \frac{-\frac{1}{\sqrt{n}}.\frac{\partial l_Y(\theta_0)}{\partial \theta}}{\frac{1}{n}.\frac{\partial^2 l_Y(\theta_0)}{\partial \theta^2}}. \qquad (18)$$

Now the top part of this fraction has mean value zero:

$$E\,\frac{-1}{\sqrt{n}}.\frac{\partial l_Y(\theta_0)}{\partial \theta} = E\,\frac{-1}{\sqrt{n}}.\sum_{i=1}^{n}\frac{\partial l_{Y_i}(\theta_0)}{\partial \theta} = \frac{-1}{\sqrt{n}}.\sum_{i=1}^{n}E\,\frac{\partial l_{Y_i}(\theta_0)}{\partial \theta} = 0$$

since θ_0 is the true value of cannonical parameter θ (see section 2.3. equation (1)).

And the formula from Theorem, section 2.3., together with previous result lead to following expression of variance

$$\mathrm{var}\left(\frac{-1}{\sqrt{n}}.\frac{\partial l_Y(\theta_0)}{\partial \theta}\right) = \frac{1}{n}.E\left(\frac{\partial l_Y(\theta_0)}{\partial \theta}\right)^2 = \frac{1}{n}.E\left(\sum_{i=1}^{n}\frac{\partial l_{Y_i}(\theta_0)}{\partial \theta}\right)^2$$

Let's define measure $\mathcal{I}$ known as *Fisher's Information* as:

$$\mathcal{I} = E\left(\sum_{i=1}^{n}\frac{\partial l_{Y_i}(\theta_0)}{\partial \theta}\right)^2$$

Note: Clearly the measure increases with the size of the dataset.

And so:

$$\mathrm{var}\left(\frac{-1}{\sqrt{n}}.\frac{\partial l_Y(\theta_0)}{\partial \theta}\right) = \frac{\mathcal{I}}{n}.$$

Using the fact, that $\frac{\partial l_{Y_i}(\theta_0)}{\partial \theta}$ for $i = (1, \ldots, n)$ are mutually independent and identically distributed random variables we can apply the Central limit theorem and get:

$$-\frac{1}{\sqrt{n}}.\frac{\partial l_{\Upsilon}(\theta_0)}{\partial \theta} \xrightarrow{D} N\left(0, \frac{\mathcal{I}}{n}\right), \text{ where } n \to \infty.$$

For the bottom part of the fraction (18) from the Law of large numbers we get:

$$\frac{1}{n}.\frac{\partial^2 l_{\Upsilon}(\theta_0)}{\partial \theta^2} \xrightarrow{P} \frac{1}{n}.E\left(\frac{\partial l_{\Upsilon}(\theta_0)}{\partial \theta}\right)^2 = \frac{\mathcal{I}}{n}, \text{ where } n \to \infty.$$

Finally, combining last two result and equation (18) we get the limiting distribution :

$$\sqrt{n}.\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} N\left(0, \frac{n}{\mathcal{I}}\right), \text{ where } n \to \infty.$$

This leads to:

$$\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} N\left(0, \mathcal{I}^{-1}\right), \text{ where } n \to \infty.$$

In establishing these results, we have considered only the case where only single parameter has to be estimated. This enabled us to proceed without using vectors and matrices. Neverheless, nothing essential has been ommited. In the case where $\boldsymbol{\theta}$ is a k-dimensional vector, we define the *Information matrix* $\boldsymbol{\mathcal{I}}$ (extension of Fisher's Information) with following elements

$$\mathcal{I}_{j,l} = E\left(\frac{\partial l_{\Upsilon}(\boldsymbol{\theta}_0, \phi)}{\partial \theta_j}.\frac{\partial l_{\Upsilon}(\boldsymbol{\theta}_0, \phi)}{\partial \theta_l}\right), \text{ where } j, l \in \{1, \ ..., k\}.$$

Therefore, in case of k-dimensional vector $\boldsymbol{\theta}$**,** the result is a vector of parameters

$$\hat{\boldsymbol{\theta}} \xrightarrow{D} N_k\left(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1}\right). \text{ where } n \to \infty.$$

This represents general result for maximum likelihood estimate. Hence, consequently for maximum likelihood esmitate $\hat{\boldsymbol{\beta}}$ in which we are interested, using equation (5) from section 2.4 it holds:

$$\hat{\boldsymbol{\beta}} \xrightarrow{D} N_k\left(\boldsymbol{\beta}, \boldsymbol{\mathcal{I}}^{-1}\right), \text{ where } n \to \infty \text{ and}$$

where the Information matrix $\boldsymbol{\mathcal{I}}$ has following elements

$$\mathcal{I}_{j,l} = E\left(\frac{\partial l_{\Upsilon}(\boldsymbol{\beta}, \phi)}{\partial \beta_j}.\frac{\partial l_{\Upsilon}(\boldsymbol{\beta}, \phi)}{\partial \beta_l}\right), \text{ where } j, l \in \{1, \ ..., k\}.$$

Note: Usually the Information matrix $\boldsymbol{\mathcal{I}}$ will not be known, until $\boldsymbol{\theta}_0$ is, and it has to be estimated by putting $\hat{\boldsymbol{\theta}}$ into the expression for $\boldsymbol{\mathcal{I}}$.