Stockholms universitet

GAMs in non-life insurance pricing

Adam Pettersson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:5
Matematisk statistik
Juni 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2020:5**
http://www.math.su.se

GAMs in non-life insurance pricing

Adam Pettersson*

June 2020

Abstract

In this thesis, the predictive ability of two types of insurance pricing models is compared. We use cross-validation on the mean squared error for measuring this. The analysis is based on actual insurance data from the nineties. The main focus is on the value of replacing linear expressions used in a classical generalized linear model with so-called smoothing splines. If the replacement leads to lower prediction errors, this will help insurance companies provide customers with more fair prices. Unfortunately, we cannot conclude anything significant. It does, however, seem like the more advanced model has a slight edge on the simpler model when it comes to modeling variables with many levels and seemingly continuous behavior.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: adam.gideon.pettersson@gmail.com. Supervisor: Felix Wahl, Mathias Millberg Lindholm, Kristofer Lindensjö.

Thanks

Ett stort tack riktar jag till Felix Wahl, för engagemang och värdefull återkoppling på tidigare utkast av arbetet. Mathias Millberg Lindholm förtjänar även han ett omnämnande, för förslag på bra källor och för data att analysera.

Contents

1 Introduction

Goal

This thesis aims to compare classical generalized linear models with generalized additive models on actual insurance data, for predicting claim amounts and claim severity.

Generalized linear models (GLM) are a class of models with a rich associated theory and are used in many applications. They date to the seventies but were used a long time before that, according to Agresti in [Agresti, p.116]. Generalized additive models (GAM) is a more recent generalization of the GLM:s, introduced in the eighties, see [Ohlsson, p. 102]. Details of similarities and differences are discussed in the theory chapter.

The aim is to compare the predictive ability of these models relative to each other. We will be predicting claim amounts and claim severities, two terms appearing in non-life insurance. The structure of the comparison is first to create GLM models, as good as possible, and then corresponding GAM models, using a large subset of our total data. Using these models, we predict the claim amounts and claim severities on a small subset of the data. The mean squared error of prediction is then calculated. That way, we measure predictive ability and avoid false conclusions caused by overfitting or insufficient measures.

Purpose

If GAM models outperform GLM models, it would be good news for insurance companies, allowing them to better price insurance contracts, hopefully resulting in lower prices for customers. The two model types are similar in many senses, which will be described in the theory section. In general, it is hard to find guidelines for when a GAM would be better than a GLM. Nevertheless, considering that Esbjörn Ohlsson added a GAM chapter to his book and described them as "powerful," in [Ohlsson, p. vii], it seems reasonable that GAM-models have some perks.

Data

We have a dataset on motorcycles. The data comes from the years 1994-1998 and is adjusted for inflation. The source is Wasa Insurance, which nowadays belongs to Länsförsäkringar. It consists of 64500 rows, each corresponding to an insurance policy. For each policy, we have several factors, as well as the number of claims and (average) claim severity. In table 1 below is an excerpt from the data.

Table 1: First 10 rows (insurance contracts) of data

O. Age	Gender	Zone	V Class	V. Age	B Class	Duration	Claim #	Severity
0	M	1	4	12	1	0.18	0	0
4	M	3	6	9	1	0.00	0	0
5	K	3	3	18	1	0.45	0	0
5	K	4	1	25	1	0.17	0	0
6	K	2	1	26	1	0.18	0	0
9	K	3	3	8	1	0.54	0	0
9	K	4	3	6	1	0.00	0	0
9	M	4	4	20	1	0.50	0	0
10	M	2	3	16	1	0.15	0	0
10	M	4	2	17	1	0.52	0	0

- **Age (of the owner).** In years, with totally 82 levels
- **Gender** Female or Male
- **Zone** Geographical zone, 7 levels
- **Vehicle-class** Type of vehicle, depending on power and weight. 7 levels
- **Age (of the vehicle)** In years, with totally 81 levels
- **Bonus-class** Grading policyholders from 1 to 7, depending on how accident-prone they have been in the past. 1 being the most likely to have an accident, 7 being the least likely.
- **Duration** How long the policy has been active. We will use these weights for the claim frequency analysis.

And the following response variables:

- **Claim Frequency** The number of claims for the policy.
- **Claim Severity** The average claim cost. When modeling this, we will condition on the number of claims, using the claim frequency as

weights.

We will see later, in the modeling chapter, how some levels of the categorical variables are joined, and some variables even removed, to find the most suitable model. A few observations are removed since their duration is zero.

2 Theory

Tariff analysis

What follows is an introduction to the terms and notations used in the book by Esbjörn Ohlsson. We will begin by looking at the underlying assumptions of the model building. The goal is to understand the terms used when pricing insurance policies. The premium itself is based on different properties. These are:

- *Properties of the policyholder:* Such as age, if a person, or type of industry, if a company.
- *Properties of the Insured object:* What type of object to be insured, how old is it, what safety precautions exist.
- *Properties of the geographical region:* Country, urban or rural.

For example, insuring a motorcycle may cost differently for an old female living in Stockholm, compared to an 18-year-old male living in Luleå. The type of motorcycle, as well as the age, will probably inflect the price. With data on previous claims, we can see claim numbers and average claim size. This allows us to compare different policies and their prices relative to each other.

The whole table of the relative difference between premiums, based on properties of the premiums, is called a tariff. In the table, age and gender are categorical variables whose combinations make up the tariff cells. Below is an explanation of the terms in a tariff.
Duration is the total amount of time the particular insurance is active for this cell
Number of claims is the total number of occurred incidents for this cell
Claim frequency Number of claims normalized with regards to Duration, in some time unit
Claim severity Is the average cost per claim
Pure premium Claim frequency times claim severity

Table 2: Dummy example of insurance data

Age	Gender	Duration	Claim number	Claim Frequency	Severity
1	1	1	3	3.0	12500
2	1	10	5	0.5	1330
3	1	5	5	1.0	992
1	2	20	100	5.0	702
2	2	50	0	0.0	0
3	2	10	2	0.2	24000

Above is the table itself. There was not space enough for the pure premium, but it is easily calculated by multiplying claim frequency with claim severity. In table 2, a row represents an insurance contract. It has been active for the time stated in the duration column. The claim number column tells how many injuries have occurred during the duration. Dividing the number of claims by the duration gives the claim frequency. The technical details of claim settlements and similar features have not been taken into concern.

In building our statistical models, we adopt some assumptions to work with. Esbjörn Ohlsson uses the following three assumptions:

1: **Independence between insurance policies**
This not entirely realistic assumption states that the outcomes of all two different policies are independent. We could see this violated if two insured persons collide with their motorcycles.

2: **Independence in time**
This means that outcomes of policies are assumed independent of previous and later outcomes. For example, we believe the number of accidents an insured person will have during a year to be independent of how many accidents the person had the previous year.

3: **Homogenenity within cells**
This means that you and your similarly-aged neighbor are assumed to be identical if buying the same vehicle.

Without these assumptions, it would be necessary to account for time-dependence and correlation between insurance policies, which likely would

be extremely messy to model. With these assumptions, we may use GLM models, a popular class of models that is easily modeled in R.

If we had enough data, we could model every cell on its own. This is hardly ever the case, so we need a way to model data despite having few observations in some cells. We achieve this using a multiplicative model. There will be some base level, and changing the policy conditions should then change the price of the policy. For example, let us say we have 7 regions, 2 age categories(young/old), and 2 vehicle categories(light/heavy). The relative price for a policy is

$$\mu_{i,j,k} = \gamma_0\gamma_{1i}\gamma_{2j}\gamma_{3k},$$

where the second factor depends on region, the third factor on age and the fourth and vehicle category. We set a base level, for example, a young person with a lightweight vehicle living in Stockholm. Then the last three factors are all set to 1. Should the person live in another region, the relative price goes up and down depending on the γ_{1i}-for that region. This is the simplest case where we assume independence between factors. This is not always the case.

The goal of the tariff analysis is mainly to determine the prices of the cells relative to each other. Later on, we will model the number of claims, but when testing the prediction error, we will focus on the number of claims between aggregated cells. The main reason for this is that accidents are rare, so even the most accident-prone people are unlikely to experience an accident in a given year. So a good predictor would be setting all cells to zero. However, such a model would not be useful when pricing insurance premiums. To still test the predictive ability, we will be looking at the cells relative to each other instead. In practice, this means aggregating over some variables and comparing the mean squared errors of prediction for the different models. Above we took an arbitrary base level, but the baseline cell is traditionally one with considerable exposure,[Agresti, p. 268].

One way to go about this problem is by using the mentioned multiplicative model, where we begin by assuming we have no interaction between our variables, i.e., that explanatory variables are independent. Then, our estimate for some cells pure premium can be expressed as a product of k+1 factors, when we have k explanatory variables. An easy way to consider interactions is by including another factor, for interaction. So if we believe age and region interact, we include a factor depending on the levels of the other two variables. This loss of independence means that changing from one region to another will cause one effect, changing from one age to another one effect, but doing

both changes may not have an effect equal to multiplying those two effects. Digging deeper into this multiplicative model isn't necessary since we will be using the more advanced GLM models, containing the basic multiplicative model described as a special case.

Cross-Validation

Later on, we will compare our models' predictive ability using cross-validation. The idea is to split the data into training and test data. The model is then built on the training data, and the mean squared error is calculated on the test data. This protects us from mistaking overfitting for having a truly good fit.

Deviance, AIC and ML-fitting

Letting $l(y, \theta)$ be the log-likelihood function of a distribution, we denote with D the deviance function. It is defined in [Agresti, p119] as,

$$D(y, \hat{\theta}_0) = -2(l(y, \hat{\theta}_0) - l(y, \hat{\theta}_s)),$$

comparing the likelihood of our data y with parameter estimates $\hat{\theta}_0$ with that of the saturated model (as many parameters as data points). Given a model the saturated model has the greatest likelihood, meaning that the deviance is always positive. The deviance is one way of measuring how well a model fits data.

Maximum likelihood-fitting builds on the idea of finding the most likely set of parameters, given a model. This is done by maximizing the likelihood function. Often this has to be done numerically. One way of doing this is, as in [Wood, p.76] by Newton-Rapson applied on the derivative of the log-likelihood, equivalent to iterating according to

$$\hat{\theta}_{k+1} = \hat{\theta}_k - f''(x_k)^{-1} f'(x_k),$$

where $f'(x_k)$ and $f''(x_k)$ are the gradient and hessian of the log-likelihood. When this converges, we will have the most likely estimates for our model.

We also note that the Akaike information criterion (AIC) for a model is given by

$$\text{AIC} = 2p - 2l(y, \hat{\theta}_0),$$

where p is the number of parameters. So the AIC increaseas if the model has many parameters and decreases with a good fit. In a comparison between

models, a low AIC is attractive[Wood, p.77]. As the models are built, later on, AIC, together with visual analytics, is used as the selection criterion.

Generalized Linear Models

There are many situations where classical linear regression is insufficient. For example, the data may not be normally distributed. Furthermore, it is sometimes beneficial to model something other than the mean of the random variables. If we believe that the logarithm of the mean can be expressed as a linear function of some explanatory variables (as in the multiplicative model), GLM offers a way to handle this. The generalization is two-folded:

· The observations belong to the exponential dispersion family, which among its special cases contains the normal distribution.

· The transformed mean is a linear function of the predictor variables

All models from this class have three components, described below

Random Component The random component of a GLM defines the probability/density function of the response variable Y_i. The distribution is assumed to belong to the exponential dispersion family. There are several variations of this. Or rather, degrees of generalization. Agresti uses a special case of Ohlsson, who in turn uses a special case of the one used in Wood. The one we will use is the one used in Esbjörn Ohlsson's book, [Ohlsson, p. 17, 2.1]:

$$f_{Y_i}(y_i, \theta_i, \phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right),$$

where Y_i is the response variable of cell i. We see that each Y has a unique parameter θ_i and one parameter ϕ which is shared among all estimators. The involved symbols are:

y_i: The i:th of our n observations.

Y_i: The response variable of the i:th observation. Assumed to follow the density function given.

θ_i: The canonical parameter.

ϕ: The dispersion parameter.

w_i: The duration/exposure. In our case, this is the duration of an insurance contract.

$b(\cdot)$: The cumulant function. Assumed to be twice continuously differentiable, with invertible second derivative.

$c(\cdot)$: A function, independent of the canonical parameter.

Systematic component The systematic component of the GLM relates a linear predictor of the explanatory variables to each subject *i*. We have for each subject i that $\eta_i = \sum_j \beta_j x_{ij}$, where the η:s connect to the random components through the link function.

Link component With the mean of the each subject being μ_i, the link function $g(\cdot)$ is such that $g(\mu_i) = \eta_i = \sum_j \beta_j x_{ij}$.

In words, we believe our data comes from a random variable Y belonging to some exponential dispersion family, and that the transformed mean can be expressed as a linear combination of some variables.

Example 1 With linear regression under the classical standard assumptions, each Y_i has the normal density function $\frac{1}{\sqrt{2pi\sigma^2}} \exp(-\frac{(y_i-\mu_i)^2}{2\sigma^2})$. We may rewrite this as $\exp(\frac{y_i\mu_i - \frac{1}{2}{\mu_i}^2 - \frac{1}{2}y_i^2}{\sigma^2} + \log(\frac{1}{\sqrt{2pi\sigma^2}})) = \exp(\frac{y_i\mu_i + b(\mu_i)}{\sigma^2} + c(y_i, \sigma^2, w_i))$, which belongs to the exponential dispersion family, with all durations equal to 1, and canonical parameter $\theta_i = \mu_i$ and dispersion parameter $\phi = \sigma^2$. The link function is the identity function, so that $\mu_i = \sum_j \beta_j x_{ij}$.

So the classical linear regression belongs to the GLM class of models. The two most relevant for this thesis are the ones with Poisson and gamma distribution. For clarity, those are derived as well, in example 2 and 3.

In our modeling, We assume the number of claims to be Poisson distributed (X_i), but we need to take the duration (w_i), i.e, how long the policy has been active, into account when modeling claim frequency $(Y_i = \frac{X_i}{w_i})$. For this purpose, Esbjörn Ohlsson defines the relative Poisson variable, [Ohlsson, p.19, 2.3].

Definition 1: Relative Poisson variable.
Let X_i be a poisson distributed variable with mean $w_i\lambda_i$, for real numbers w_i, λ_i. Then $Y_i = \frac{X_i}{w_i}$ has frequency function(for y_i such that $w_i y_i$ is an

integer):

$$e^{-w_i\mu_i}\frac{w_i{\mu_i}^{w_iy_i}}{(w_iy_i)!}$$

and is said to be relative poisson distributed. In our case X_i is the number of claims and Y_i the claim frequency.

We continue by noting that the relative Poisson variable belong the the exponential dispersion family.

Example 2 Rewriting the frequency function in definition 1 as

$$\begin{aligned} e^{-w_i\mu_i}\frac{w_i{\mu_i}^{w_iy_i}}{(w_iy_i)!} &= exp(-w_i\mu_i + w_iy_i \cdot log(w_i\mu_i) - log(w_iy_i!)) \\ &= exp(\frac{(y_ilog(\mu_i) - \mu_i)}{1/w_i} + y_iw_ilog(w_i) - log(w_iy_i!)) \\ &= exp(\frac{(y_ilog(\mu_i) - \mu_i)}{1/w_i} + c(w_i, y_i)) \end{aligned}$$

we see that it indeed belongs to the GLM-family, with $\theta_i = log(\mu_i), \phi = 1$.

Example 3 The claim severity is traditionally modeled as a weighted gamma distribution, where the weights are the number of claims[Ohlsson, p.20]. This means that if our response is y_i we believe this to be the mean of w_i different gamma distributions (representing injury claims) with parameters α, β. More formally, we model

$$Y_i = \frac{Y_{i1} + .. + Y_{iw_i}}{w_i} \in Ga(w_i\alpha, w_i\beta),$$

where the last step follows due to properties of the gamma distribution. The gamma function belongs to the dispersed exponential distribution, making it a GLM.

When we later model using GAM:s, we will use these same distributions. The only thing that changes is the systematic component.

This theory is relatively recently developed. The most common models had been known before[Agresti, p.116], but tying the theory together makes it possible to generalize, for example, the ML-estimation. So the ML-estimation

is done on the general case, and for a specific model, we simply input the parameters and have a method for ML-estimation on any GLM-model. The technicalities behind fitting a GLM-model are in the appendix.

Splines

Splines are functions common when interpolating and smoothing. They consist of piecewise polynomials. Given an interval we wish to interpolate or smooth, this interval is divided into subintervals. Let us for formality say that we want a spline over the interval [a,b]. Consider

$$a = x_0 \leq x_1, ... \leq x_n = b,$$

so that the union of the subintervals

$$[x_0, x_1], [x_2, x_3], ..., [x_{n-1}, x_n]$$

is exactly [a,b]. Further, for each of the k intervals, we define a polynomial P_i on the interval $[x_i, x_{i+1}]$. The spline is then the function taking values $P_0(x)$ when x is in $[x_0, x_1]$, $P_1(x)$ when x is in $[x_1, x_2]$ and so on, until $P_{k-1}(x)$ when x is in $[x_{k1}, x_k]$. The points $x_0, x_1, .., x_k$ are called the knots of the spline. In figure 1 below is an example of a cubic spline(in red) fitted to some data. The points represent observations and the red line is the spline.

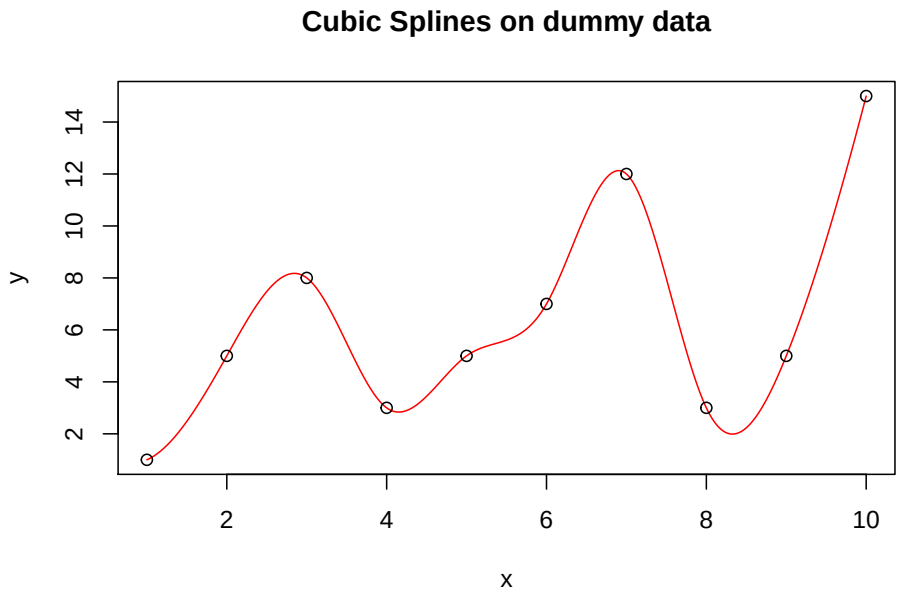

Figure 1: A cubic spline (in red) fitted to some data points

The spline has to satisfy some restrictions, such as having continuous values and continuous derivatives. In plot 2.1 above, we had a cubic spline, which must satisfy being twice continuously differentiable over the interval. We had ten observations, and they were all used as knots, implying that the spline passed through all the observed points. This gives a perfect fit and deviance equal to zero. However, in practice, when using GAM:s, we turn to penalized smoothing splines. Then we penalize with regards to the smoothness of the spline, with smoothness defined as

$$\int_a^b (f''(x))^2 dx.$$

This is illustrated and more formally described in the next section. In particular, equation (1) on page 15 shows the balance between the goodness of fit and smoothness.

Returning to the creating of splines, note that the spline passing through the observed points is not a necessary condition. To achieve more smoothness, we have to abstain from some of the goodness of fit. In the package we will use later, the k knots are chosen so that they have an equal number of points between them, as described by Wood in [5]. Once the k points are chosen, the cubic spline is selected based on a combination of a good fit to the data and wiggliness. To actually determine the coefficients of the piecewise polynomials would require us to dig deeper into the theory which is given in [Ohlsson, p. 108]. Instead of doing that, the mgcv R package is used for modeling with splines. The package is described in the appendix.

Generalized additive models.

Some variables have a huge range of possible values. One approach we will use when modeling such variables is categorizing the data. In the best of worlds, the observations in the same category are similar and, hopefully, we have enough data to properly estimate the parameters. This is not always the case. Furthermore, even if possible, finding suitable categories can be a time-consuming job. For this purpose, the Generalized additive models are introduced. The idea is to fit continuous functions for the variables where we feel that categorizing is not sufficient. There are many ways to do this. We will restrict ourselves to cubic splines, which turn out to satisfy some attractive properties. In practice, this means replacing the systematic component of the GLM with a similar, more general expression. We recall

from previously the look of the systematic component,

$$\eta_i = \sum_j \beta_j x_{ij}.$$

With the GAM generalization, this becomes

$$\eta_i = \beta_0 + f_1(x_{i1}) + ... + f_J(x_{iJ}).$$

In our data, modeling with splines will be of interest only on the vehicle age and age parameters, which will render a model like, in the case where we do not account for interaction terms:

$$\eta_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \sum_{j=3}^{J} \beta_j x_{ij}.$$

Once we do account for the interaction between gender and age, we will do this by having two separate functions, where gender decides which function to use. To find the functions, we need some restrictions. We have at our hands the observed values $\{x_{1j}\}$, to which we want to find a smooth function f_1. Smoothness means here that it is two times continuously differentiable.

Smoothness is not all that matters. It is necessary to have a good fit to data. Needing a measure taking into regard both the deviance and the smoothness of the function, a suitable option is, [Ohlsson, p. 104, eq. 5.4]:

$$\Delta(f_1) = D(y_1, \mu_1) + \lambda \int_a^b (f''(x))^2 dx. \tag{1}$$

The first term is the deviance, while the second term measures the variability of the function f. The λ decides how we balance between small variability and a good fit. For a given model, in figure 2 below is shown how the result depends on λ. We see how a growing λ fits the data less and less, but instead, become smoother and smoother.

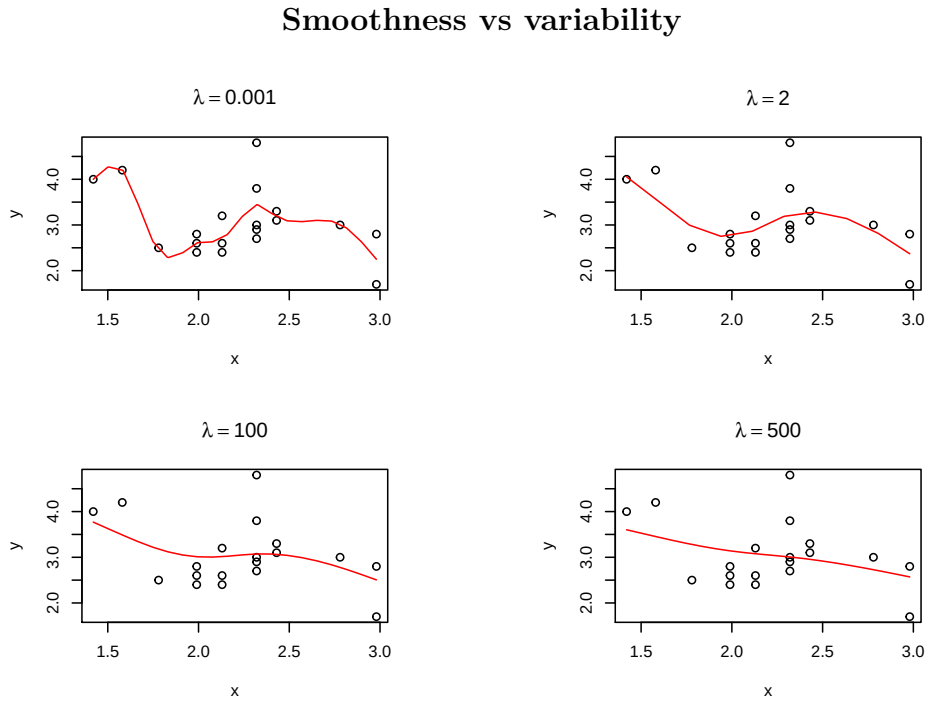

Figure 2: Cubic splines fitted to data, with varying degrees of smoothness, quantified by smoothing parameter λ. The plots were made using code found in [Wood, p. 169].

But we need a value for λ. There are several ways of doing this, and none is very easy. The conceptually easiest method, which according to [Wood, p. 269] may also be the best, is slow, but works on the idea of applying Newton's method to find the optimal λ. This requires computing the model for each iteration, and a measure (Generalized approximate cross-validation in our case, which is out of the scope of this thesis) on the model for each λ, to determine when to convergence the is sufficient.

3 Model building

In this chapter, we will start by looking at claim frequency. We will analyze the data and build two types of GLM:s. In the first one, we create levels for the variables using visual inspection, and then remove some levels and even entire variables, with AIC as a measure of model strength. In the second model, we proceed similarly. However, the variables with a massive number of levels, age of the owner, and age of the vehicle, are treated differently.

They are divided into about 20 levels each, with each level corresponding to a 5-year interval. These two models are then compared to a GAM, which is identical, apart from that the owner age and vehicle age, as well as the interaction between owner age and gender, are modeled with cubic splines.

A similar approach is used for claim severity, with some minor differences in which variables to include. We will be looking at plots and tables to compare and try to understand the results. Our model building is not perfectly executed here, but that is not very important since we are interested in the effect of the GAM difference and not optimization of the total prediction.

Claim Frequency

Modeling

If we include all variables in the original format with all variables, the number of cells is vast. In particular, the owner's age and the vehicle age have many levels. This means that if we proceed with the data in this form, the number of parameters will be large. While that will give a good fit to the data, the parameter estimates will be insecure, and the good fit likely the result of overfitting. Our first step is thus to eliminate the number of parameters. Visual analysis of the data can be of help here, as well as AIC. To get started, I will divide the two largest categories into approximately equally sized smaller levels. Then I will analyze the AIC and see if I can fuse some levels. The reason for preferring fewer levels and fewer variables is that the standard error of our estimates becomes smaller, making sure we include less noise.

Below, in figure 3, is a graphical analysis before we start modeling the claim frequency.

Graphical inspection of data

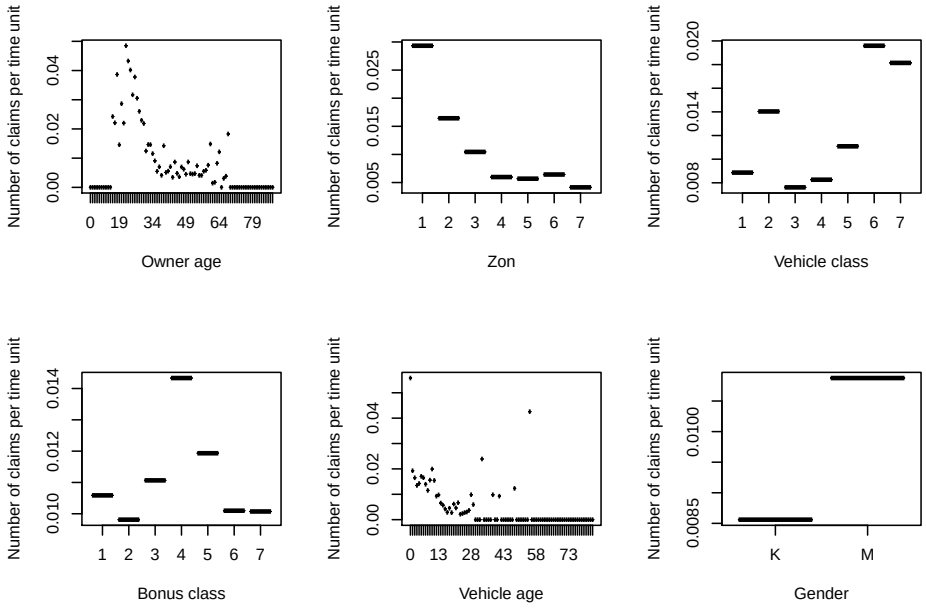

Figure 3: Distribution of claim frequency over the different levels, for each of the variabels. We can see that young people have higher claim frequency than old people (plot 1), and that men have higher claim frequency than women (plot 6).

In the top middle graph, the data hints about Zone 4,5,6 being similar. Fusing the levels into one resulted in a lower AIC. Similarly, bonus class levels 6 and 7 look similar, and also there the AIC was lowered by fusing the levels. A few more of these AIC-lowering merges were made. On top of that, the AIC of submodels was tested, by leaving one variable out at a time, and then comparing AIC. It turned out the Bonus Class was either confounding with some variable or plainly unnecessary. Either way, it was left out. This part of the modeling was not very structured, which is forgivable since our primary goal was not to find an optimal model but to compare GAM and GLM models.

We are now ready to model GLM and GAMs for claim numbers. Before proceeding, the data is split into test and training data. The training data is a randomized sample containing 80 percent of the original observations, and the remaining observations are in the test set.

Claim frequency results

With the models at hand, MSEP was calculated and is presented in Table 3 below.

Table 3: Mean Squared Error of Prediction

GLM	GAM	GLM 2
132.17	131.27	131.63

Where GLM is

$$log(\mu_i) = log(w_i) + \sum_{j=0}^{6} \beta_j x_{ij} \ + \beta_{25} x_{i2} x_{i5},$$

with x_{i0} being equal to 1 for all i:s, making β_0 the intercept. All variables are categorical, coded using dummy variables. Coding in dummy variables means that the parameter β_j and corresponding x_{ij} for observation i, are vectors. If the numbers of levels of the categorical variable x_{ij} is N, these vector have length N-1. Let us say that N = 6. Then observation i belonging to category 1 in the j variable corresponds to $x_{ij} = (0, 0, 0, 0, 0)$. Category 2 corresponds to $x_{ij} = (1, 0, 0, 0, 0)$, category 3 to $x_{ij} = (0, 1, 0, 0, 0)$,.. and category N (i.e. 6) to $x_{ij} = (0, 0, 0, 0, 1)$. The elements of β_j vector are what we estimate in this model.

GLM 2 is

$$log(\mu_i) = log(w_i) + \sum_{j=0}^{6} \beta_i x_{ij} \ + \beta_{25} x_{i2} x_{i5}.$$

It is identical in model type and variables included, but the levels are different. Specifically, what separates the two GLM models are the different levels of the age and vehicle age variables.

GAM is

$$log(\mu_i) = log(w_i) + \sum_{j=0}^{4} \beta_i x_{ij} + f(x_{i5}) + f(x_{i6}) + f(x_{i2}, x_{i5}),$$

so that we have replaced the linear terms for vehicle age and owner age with smooth functions, and the interaction between age and gender with a smooth function. In practice, this corresponds to having two smoothing splines for

owner age and letting the gender determine which one to use. This means we could equally well write the model without the $f(x_{i5})$. For a comparison between the models, see the table 4 below.

Table 4: Comparison between models

Name	Code	GLM	GLM_2	GAM
Gender	xi2	2 levels	2 levels	2 levels
Zone	xi3	4 levels	4 levels	4 levels
V. Class	xi4	3 levels	3 levels	3 levels
O. age	xi5	4 levels	16 levels	Continuous
V. age	xi6	3 levels	15 levels	Continuous

This result in table 3 is not a very good one, and not very telling either. A model setting all claims to zero would likely have been better at predicting claim numbers. The reason for this is that injuries are so rare. So we need some other way of evaluating the GAM model. One idea that comes to mind is looking at the aggregated variables for each involved variable and then look at that MSEP. This way, we will have fewer cells and thus more injuries in each cell. In the table below, the MSEP is displayed for each variable separately. For example, with the gender variable, the value was calculated by aggregating the total number of predicted claims over males and women and then calculate the mean squared error of this prediction for each model. So the value displayed in table 5 is the sum

$$(\text{predicted male injuries- actual male injuries})^2+$$

$$(\text{predicted female injuries- actual female injuries})^2.$$

Similarly, the MSEP for owner age is calculated as a sum of 82 terms, one for each age represented in the data. The difference between these and the values in table 3 is that table three is the sum of several thousands of observations, one for each cell. Here, in table 5, we have the same cell predictions as in table 3, but we have aggregated over variables.

Table 5: MSEP over aggregated categories

	GLM 1	GAM	GLM 2
Owner Age	227	175	208
Zone	355	364	382

	GLM 1	GAM	GLM 2
Vehicle Class	230	233	239
Vehicle Age	205	135	175
Gender	382	401	417

These results indicate that the easiest models work slightly better than the GAM at the variables with few levels, but that the GAM works better at owner age and vehicle age. At the vehicle age, the GAM is far better. To possibly understand these results, we look at the spline created for vehicle age, and the splines for age on men and age on women.

In figure 4 we compare this to the corresponding estimates made by the GLM:s.

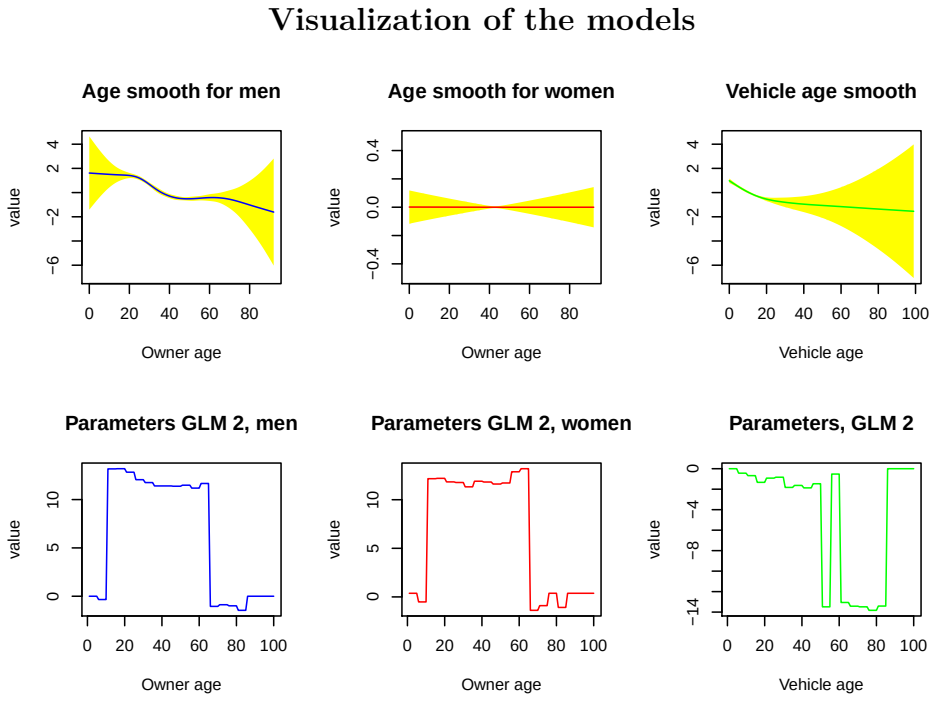

Figure 4: Above are GAM splines followed by GLM parameter estimates underneath. The splines are surrounded by yellow confidence bands. These are constructed with standard errors. For each point on the spline, a point is added two standard errors above and two standard errors below the point[Wood 6]. They hint about the preciseness of the splines.

We see that the scales of the splines are significantly smaller than the corresponding GLM estimates. A probable explanation lies in the different intercepts for the GAM model and the GLM 2 model. The difference is about 11.5, which, together with the penalization of the splines (preventing it from diving so radically at the higher ages), could perhaps explain this difference. Nonetheless, this scale difference is not something we will examine further. Apart from the scale difference, the owner age plots for GLM 2 seems to peak unreasonably early, considering the age limit for motor vehicles in Sweden. The explanation for this is that most ages (even toddlers) have some insurance contracts, but not all (see table 1). So the beginning and the end of the x-axis are improper. This error was discovered late and will not be corrected.

Disregarding that, we can see that the overall trend is similar between the splines and the GLM 2 estimates of owner age. It peaks around twenty and then falls slowly.

This is hard to interpret, but we can notice the difference between the Vehicle effect in GLM2 and GAM. The GAM has a smooth looking curve, while the GLM 2 is bumpy. The main thing distinguishing the Vehicle age is the sparse number of observations in higher age categories. Considering how the GAM predicted the vehicle age better, it seems like the vehicle age benefits from the penalization. It thus seems like vehicle age has a high variation even within categories.

Claim severity

Moving on the claim severity, the procedure will be identical to that of claim frequency. We start by looking at data, followed by modeling and then compare the models.

Modeling

In the same way as in figure 3, we look at the data. This time, in figure 5, the y-axis represents claim severity rather than claim numbers.

Graphical inspection of data

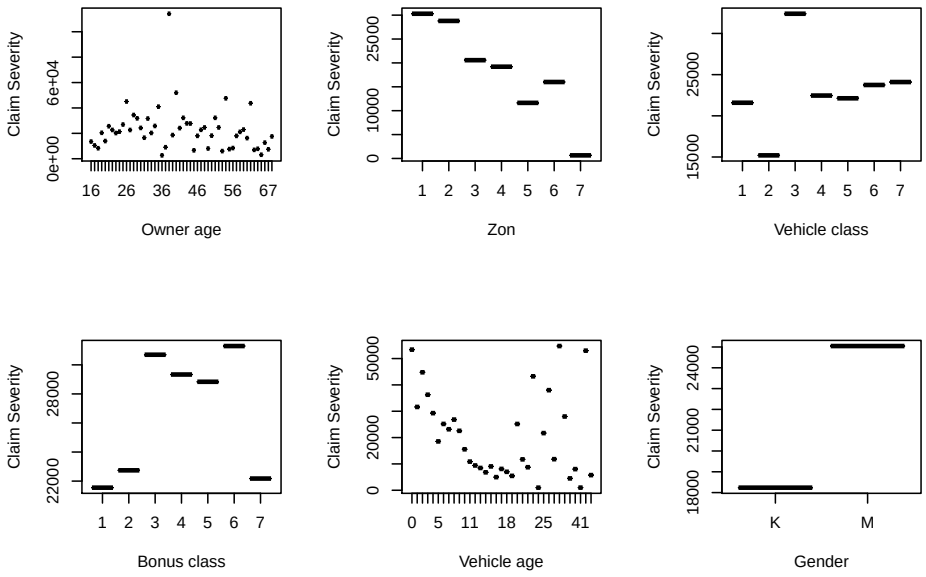

Figure 5: Distribution of claim severity over the different levels.

In figure 5, we can see that new and really old vehicles people have higher claim severity than old but not vintage vehicles. Furthermore, the age seems to have no effect. After more formal testing using AIC, models were built. Most notably, the owner age variable had no effect and was left out.

Claim severity results

In table 6 below, we again look at the MSEP. This time for claim severity instead of claim numbers.

Table 6: Mean Squared error of Prediction, claim severity

GLM	GAM	GLM 2
212702364795	207525252439	207431300908

To get a better overview, the values relative to each other are compared in table 7.

Table 7: Relative values of MSEP

GLM	GAM	GLM 2
1.02	1	1

In the previous section, we did not put much value on the total MSEP. Now, computing MSEP makes more sense. Rather disappointingly, none of the models seemed significantly better than any of the other.

We proceed by looking at some aggregated cases, as in the claim frequency section.

Table 8: MSEP over aggregated categories. For each variable, the model with the lowest MSEP is given value 0, and the other are expressed in PPM:s over this value. (To increase readability)

	GLM 1	GAM	GLM 2
Owner Age	117924	25397	0
Zone	0	1634133	1067152
Vehicle Class	0	230138	22672
Vehicle Age	84092	32319	0
Gender	0	685496	548710

In table 8 above, we see how the GAM model fails to model any variable the best. Recalling that GAM performed excellently at Vehicle age in the claim frequency case, this is unexpected. To hopefully gain an understanding of why we look at the vehicle age spline in figure 6 below.

Visualization of GAM and GLM 2

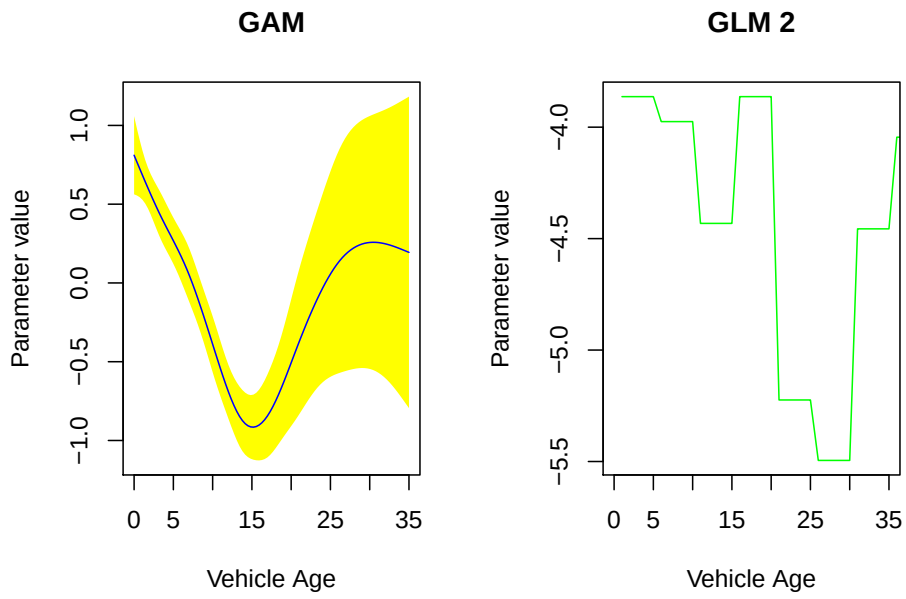

Figure 6: The spline from the game model and the corresponding parameter estimates from the GLM 2 model. r

In figure 6, the spline resembles plot 5 in figure 5, which is reasonable. The scale between the graphs in figure 6 is different, but the curves have similar ups and downs. The GLM models take a dive at the end. While the spline does not follow that dive, the yellow area is large, implying a large standard error.

What we have compared here is the claim frequency and claim severity separately. In practice, we would perhaps be the most interested in the pure premium. That means that we should use the predicted claim numbers as weights rather than the given data. Since our main focus is to experiment with the GAM model, the calculation of pure premiums is omitted.

4 Results

Claim numbers

Next, we compare relative values for the other variables as well, by aggregating. This time, however, we consider the test data and not the training data, as we did above.

Table 9: Predicted number of claims

Actual number of claims	122
GLM-prediction	145.29
GAM-prediction	145.67
GLM2-prediction	146.13

To begin with, we note in table 9 that both models significantly overestimate the true number of claims, and GAM is worse in this regard. This problem is probably not something any model could do much about; it is just the result of randomness.

Claim severity

Table 10: Predicted total cost of injuries (conditioned on the true number)

Actual total cost	3363204
GLM-prediction	3407042
GAM-prediction	3320240
GLM2	3343553

With the sum of costs of all claims, displayed in table 10, all three models perform about as good as the others.

5 Discussion

The Mean Square Error used was weak, and we were not able to conclude anything significant or make any precise statements. More efforts could have been made to find a better GAM—different kinds of training points, or perhaps different degrees of penalization. We could have looked at relativities as well, which is the industrial standard. However, with our disappointing results, it did not seem worthwhile. The GAM showed promise on the Claim Severity but did not impress on Claim numbers. With that, examining pure premium did not feel purposeful since they depend on Claim Severity. It would, though, have made the work more complete and realistic.

On the other hand, considering that the models except for one case are pure

multiplicative, we could perhaps have gotten better results by more carefully looking at each variable. Also, we did not have use for such rich data. Since the goals were to examine the value of GAM and penalization and smoothing, it could have sufficed with aggregating over age and only looking at that. We did, however, see some interesting trends by using many variables, that the GAM performed well on complicated variables, but failed hard on those with few levels. Kind of like golfer who delivers beautiful hole in ones mixed with flunking easy puts.

Simulation

Judging by these results, it seems like a scenario where GAM potentially is significantly better is when, for some variable in the training data, it is dense in some interval and non-dense in some intervals. For example, we have much data for when the variable is equal to certain values, but for some values, almost no data, throughout the data. A dummy example regarding motorcycles could be if we believe different years had different trends and that some years were inspected, and some were not. For example, maybe some federal institute examines motorcycles of a certain brand thoroughly every five years, and lightly otherwise. Perhaps they look at 20 brand new motorcycles every five years, and on other years only three motorcycles. This would probably cause a regular GLM model focusing on deviance overfitted, and instead benefit from penalized interpolation (since there is an underlying trend), as in GAM.

Or, we might believe that damage from storm claims get more and more severe the more up north in Sweden you live. We would then have a lot of data from urban areas such as Skövde and Stockholm, but in other parts of the country, the number of people living there is significantly smaller. That could give rise to a scenario where we have an underlying trend, rich data for some values on the explanatory variable, and almost no data on other values of the explanatory variable.

To construct this example, let us say we have the simple model $log(\mu : i) = \alpha + f(x_i)$ and compare this to $log(\mu : i) = \alpha + \beta x_i$, y is gamma. x range from 1 to 100. To get y:s, we simulate 100 sets of gamma parameters. There should be some trend for the models to catch up on. To make this as simple as possible, I will simulate 100 sets of parameters, order them after mean, and simply say that's the order. We will consider these 100 parameter sets to be the true values. For each set of parameters, there is a corresponding gamma distribution with some mean. This mean is displayed in red, in the

plot below. From these, we will simulate samples of observations for each of the 100 sets. The mean of the samples are the dots in the plot below.

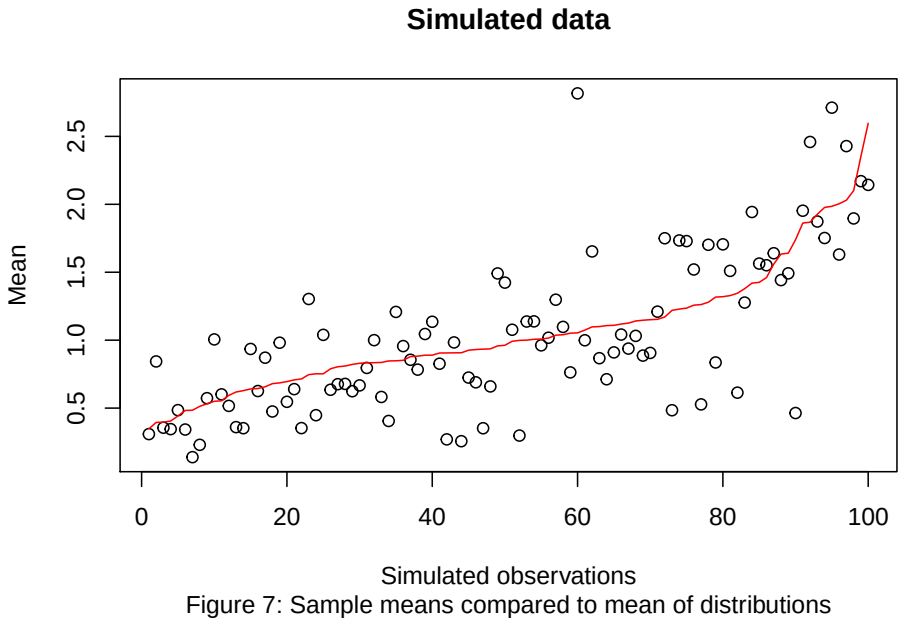

Figure 7: Sample means compared to mean of distributions

In the plot above, we have observed the means of values simulated from a gamma distribution with mean according to the red line. So at x-value 37, we have simulated from the gamma distribution with the 37:th smallest mean (whose value is given by the red line at x=37), and the dot is the mean of these simulated values. The red line is thus by definition increasing, while some earlier dots have larger values than later dots, due to the nature of the simulation.

Simulated values compared with true means

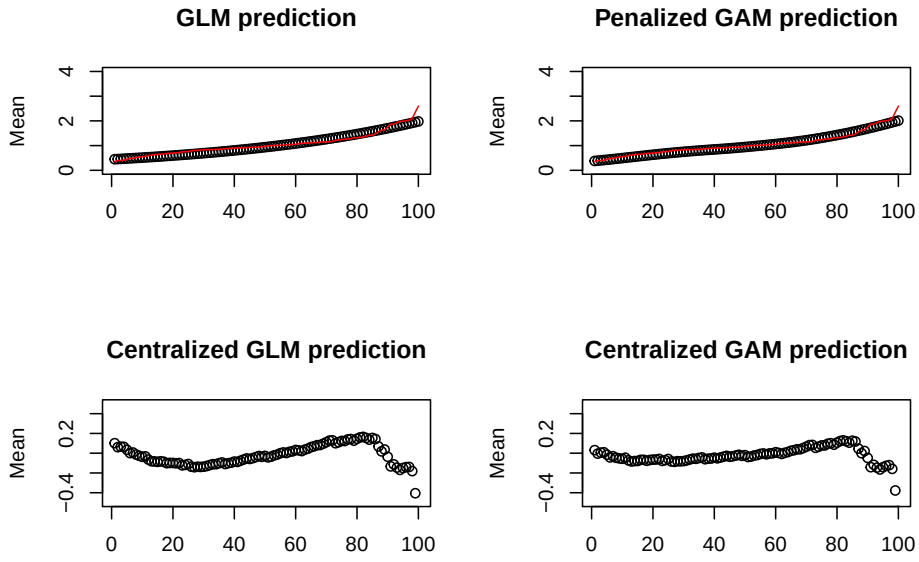

Figure 8: The predicted values compared to the means. In the centralized plots, the means (displayed in red in the upper plots) has been subtracted from the predictions.

In the upper plots of figure 8, we see how the predictions compare to the means(in red). The lower plots show the same predictions around 0, to better display the structure of the errors. We see how the errors for both models behave similarly, but that the GAM errors are smaller. Now we try to measure these errors, by calculating the true mean squared error of prediction. This means subtracting the gamma means from our predicted means and taking the sum of these squares.

Table 11: Mean squared error for the models, with simulated data

Object	MSE
GLM-model	1.484
GAM-model with penalization	0.985
GAM-model without penalization	1.115

As can be seen in table 6, the GAM was much better. One might argue

that this is not the most realistic simulation, which it is not, but at least it demonstrated how penalized regression can greatly reduce errors. One could further argue about how it exists penalized GLM models which are not GLM:s and that this is not very revolutionary, but we have not examined that in this thesis, so result is still exciting.

Conclusion

My conclusion is that GAM can work well on advanced variables, but require careful analysis, since fitting them require more computer resources. In the insurance industry, this could perhaps account for computing relativities for suitable variables separately (if independence is plausible), and then include them in the final tariff. This way, we will get the GLM model's reliability on simple variables, and GAM flexibility on the more advanced ones.

6 Appendix

Fitting a GLM-model

For given data and a chosen model, we need a way to estimate the parameters. The ML-estimates have attractive properties and are traditionally used. It can rarely be done by exact methods. One rather uses numerical methods, such as the Newton-Rapson method. We will do that below, using the exact same approach as in Wood, [Wood, p.105]. Let us assume we have an N-dimensional vector Y of independent responses, where each element Y_i belongs to the dispersed exponential family with density function $f_{\theta_i}(y_i)$. We have the mean vector μ, defined so that $\mu = E[Y]$ We believe that

$$g(\mu_i) = X_i\beta.$$

The indepence between respones implies the likelihood function is given by

$$L(\beta|y) = \prod_{i=1}^{n} f_{\theta_i}(y_i),$$

giving a log-likelihood of

$$l(\beta) = \sum_{i=1}^{n} log(f_{\theta_i}(y_i)) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i) \right).$$

For Newton-Rapson we need the gradient and the hessian.

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} w_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i) \frac{\partial \theta_i}{\partial \beta_j} \right),$$

repeated chain rule use gives

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{d\theta_i}{d\mu_i}\frac{\partial \mu_i}{\partial \beta_j} = \frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Now, remembering that $\eta_i = g(\mu_i)$ and that $g(\cdot)$ satisfies certain regularity conditions, it follows that $g^{-1}(\eta) = \mu_i$ has derivative $\frac{1}{g'(g^{-1}(\eta_i))} = \frac{1}{g'(\mu_i)}$, and similarly, since $\mu = E[Y] = b'(\theta)$, $\frac{d\theta_i}{d\mu_i} = \frac{1}{b''(\theta_i)}$. By the definition of X_i,

$$\frac{\partial \eta_i}{\partial \beta_j} = X_{ij}.$$

This together gives

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{X_{ij}}{g'(\mu_i b''(\theta_i))}.$$

Now this inserted to the original expression gives

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} \frac{y_i - b_i'(\theta_i)}{g'(\mu_i)\, b_i''(\theta_i)/w_i} X_{ij}.$$

More tedious calculations give that the second derivative becomes

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\frac{1}{\phi}\sum_{i=1}^{n} \frac{X_{ik}X_{ij}\alpha(\mu_i)}{g'(\mu_i)^2\, b_i''(\theta_i)/w_i},$$

with

$$\alpha(\mu_i) = 1 + (y_i - \mu_i)\left(\frac{d\ b_i''(\theta_i)/w_i}{d\mu_i}/(b_i''(\theta_i)/w_i) + \frac{g''(\mu_i)}{g'(\mu_i)}\right).$$

To make this more compact, define a diagnoal matrix V = diag(v_i) where $v_i = \frac{w_i\alpha(\mu_i)}{g'(\mu_i)b_i''(\theta_i)}$, allowing us to rewrite the hessian as $\frac{-X^TVX}{\phi}$. Finally, setting G = diag$(g'(\mu_i)/\alpha(\mu_i))$, gives the gradient of the log likelihood as $X^TVG(y-\mu)/\phi$. The multivariate Newton Raphson iteration now becomes:

$$\beta^{[k+1]} = \beta^{[k]} + (X^TVX)^{-1}X^TVG(y-\mu)$$

$$= (X^TVX)^{-1}X^TV\left(G(y-\mu) + X\beta^{[k]}\right).$$

Now, setting the expression in the right brackets to a vector z, where then

$$z = \left(G(y-\mu) + X\beta^{[k]}\right) \Rightarrow z_i = g'(\mu_i)(y_i - \mu_i)/\alpha(\mu_i) + \eta_i.$$

The Newton-Raphson iteration step now becomes $= (X^T V X)^{-1} X^T V z$, we can be recognized as the solution to the problem of finding the weighted least squares estimator of beta by minimizing

$$\sum_{i=1}^{n} w_i (z_i - X_i \beta)^2.$$

This allows us to find the ML-estimates using an iterative re-weighted least square algorithm, and it goes:

Starting with

$$\hat{\mu}_i = y_i, \ \hat{\eta}_i = g(\hat{\mu}_i),$$

the algorithm proceed with
1. Compute

$$z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha(\hat{\mu}_i) + \hat{\eta}_i, \ w_i(\hat{\boldsymbol{\beta}}) = w_i \cdot \alpha(\hat{\mu}_i)/(g'(\hat{\mu}_i)^2 b(\hat{\mu})'').$$

2. Minimize the weighted least square sum with regards to β, then update the vectors

$$\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \text{ and the ML-estimates } \hat{\mu}_i = g(\hat{\eta}_i)^{-1}.$$

until the change in deviance is sufficiently close to zero, or some similar test.

Fitting the GAM given lambda

Fitting a GAM is very similar to fitting a GLM model. The difference is the last step of the iteration, where we instead of just minimize a weighted least square, minimize [Wood, p. 251]

$$\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|_W^2 + \sum_j \lambda_j \boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}.$$

This looks a little simpler than it is, X and β are extended here, to account for the smooth factors as well. So in principle, the idea is the same, but performing it becomes more advanced and is skipped here.

Estimating lambda

According to [Wood, p. 269], the most reliable method for estimating λ is to use Newton-Raphson. This requires, for each outer iteration, a penalized weighted reiterated least squares iteration to find the estimates corresponding to the iteration. This means we need the hessian and gradient the expression we minimize. As one might imagine, this is not a little cumbersome, so this procedure is left out.

The mgcv package

The mgcv (Mixed GAM Computation Vehicle) package is built by Simon Wood and was released in conjunction with his book [3]. Simon Wood is a professor in Statistics at the University of Bristol and has strong interest in GAM models. He has an h-index of 53. The package is availaibe on CRAN. It contains the GAM function used for building our GAM models in chapter 3, as well as the functionality to plot the splines afterward.

7 References

[1] Agresti, (2 edition, 2001) "Categorical Data analysis"

[2] Ohlsson, Johansson (1 edition, 2010), "Non-Life Insurance Pricing with Generalized Linear Models"

[3] Wood, (2 edition, 2017), "Generalized Additive Models"

[4].Wood, https://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/smooth.construct.cr.smooth.spec.html.

[5] Wood MGCV package pdf. See, in particular, "se" under the plot.GAM section