# Nonnested model comparison of GLM and GAM count regression models for life insurance data

Claudia Czado, Julia Pfettner, Susanne Gschlößl, Frank Schiller

December 8, 2009

**Abstract**

Pricing and product development of life insurance contracts require the determination of company specific risk factors and their resulting risk rates. The paper shows how to use generalized linear models (GLM) and generalized additive models (GAM) to quantify the effect of risk factors by allowing for non linear and interaction effects. Nonnested model comparison between GLM and GAM based specifications are facilitated using non-randomized probability integral transforms (see Czado, Gneiting, and Held (2009)) and proper scores (see Gneiting and Raftery (2007)) developed for count responses. These allow for the assessment of model fit and predictive capability of a model. For a life insurance portfolio it is shown that the computationally less demanding GLM specification performs similarly to a GAM specification.

**Keywords:** Count regression, GLM, GAM, prediction, probability integral transform, proper scores

## 1 Introduction

Pricing and product development in life insurance is based on best estimate rates for risks like mortality, disability incidence and termination or lapse. Official tables provided by actuarial associations may be used as a basis. However, it is well known, that rates might differ significantly between companies. The expected claims heavily depend on the portfolio structure of a company, i.e. on different target groups, products offered etc. Hence, best estimate rates should ideally be derived by analyzing company specific portfolio data. In order to determine risk adequate premiums and avoid antiselection it is important to identify and quantify significant risk factors. Statistical models like generalized linear models (GLM's) (see Nelder and McCullagh (1989)) or generalized additive models (GAM's) (see Wood (2006)) provide a

method to do so.
The main risk driver for mortality or disability incidence rates is the age of the insured. The rates typically depend on age in a non linear way. For an adequate pricing it is essential to get the shape of the dependence on age right. In this paper, a non parametric approach is taken to model the functional form of age. We consider both a GLM where age is estimated non parametrically in advance and then entered as covariate in the model and a GAM where age is modeled non parametrically together with the remaining covariates.
A focus of this paper is model comparison. Since GLM's and GAM's are not nested, classical criteria for comparing models like the AIC or the likelihood ratio test are not efficient. Instead we consider criteria based on the predictive distribution of models, in particular the probability integral transform and proper scoring rules. The probability integral transform for count variables (see Czado, Gneiting, and Held (2009)) assesses calibration and sharpness of count models, while proper scoring can be used to compare non nested models (see Gneiting and Raftery (2007)). An application to a data set from life insurance is given.

# 2 Statistical regression models for death rates

Generalized linear models (GLM's) provide a well known class of statistical models for analyzing dependencies between a possibly non normal response variable and a number of covariates, see for example Nelder and McCullagh (1989) for details. An extension of GLM's which allows for the incorporation of non parametric covariate effects is given by generalized additive models (GAM). See for example Hastie and Tibshirani (1990) for an early account of GAM models and Wood (2006) for a later one.
Assume a data set with $n$ observations and let $\mathbf{Y} = (Y_1, ..., Y_n)'$ denote the vector of response variables. Further, $\mathbf{x_i} = (x_{i1}, ..., x_{ip})', i = 1, ...n$ denotes the vector of covariates and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ the corresponding vector of unknown parameters.

## 2.1 Generalized linear models

In a GLM, the response variables $Y_i$ are assumed to be independent given the covariates $\mathbf{x_i}$ for $i = 1, ..., n$, and to follow a distribution from the exponential family (see Nelder and McCullagh (1989)). The exponential family includes for example the Normal, Binomial, Poisson and Gamma distribution. Covariate information is incorporated by modeling a transformation of the mean in terms of covariates. For life insurance data, the response variable is typically the number of events like death, disability or lapse, i.e. a count variable. We assume a Poisson distribution for the response which is a common choice for count data. The logarithm is taken as

link function $g(\cdot)$, leading to a multiplicative model which is easy to interpret and reasonable for actuarial applications. The model reads as follows:

$$Y_i \sim Poi(E_i\lambda_i) \qquad i = 1, ..., n$$

with $g(\lambda_i) = ln(\lambda_i) = \eta_i := \mathbf{x}_i^T\boldsymbol{\beta}$. The quantity $\eta_i$ is called the linear predictor for observation $i$.
Since the number of events in life insurance is proportional to the exposure to risk $E_i$, the logarithm of the known exposure is included as an offset, i.e.

$$\mu_i := E(Y_i) = E_i\lambda_i = E_i \exp(\mathbf{x}_i^T\boldsymbol{\beta}) = \exp(\log E_i + \mathbf{x}_i^T\boldsymbol{\beta}).$$

For an account of GLM's from an actuarial science view point see also de Jong and Heller (2008).

## 2.2 Generalized additive models

An extension to GLM's is given by generalized additive models (GAM's), see Wood (2006) for an overview. The basic setting for a GAM is the same as for a GLM. In contrast to GLM's however, GAM's allow for the incorporation of non-parametric functions of continuous covariates in the linear predictor. The linear predictor $\eta_i$ in a GAM can be written as

$$\eta_i = f_1(x_{1i}) + ... + f_J(x_{Ji}) + \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}(x_{ij}, x_{ik}) + \mathbf{z_i^t}\boldsymbol{\gamma}$$

where $f_j(x), j = 1, 2, ..., J$ denote smooth functions depending on covariate $x$ which are not further specified and are estimated from the data using smoothing splines. Interactions with continuous variables can be modeled using functional terms $f_{jk}(x_{ij}, x_{ik})$. If an interaction between a categorical and a continuous variable is included, this implies the estimation of a separate smooth function of the continuous variable for each level of the categorical variable. An interaction between age and gender for example, induces the estimation of a separate functional dependency on age for males and females. For interactions between two continuous covariates two dimensional surface functions are allowed.
Finally we allow the inclusion of parametric effects. For this let $\mathbf{z_i}$ denote the vector of covariates for any strictly parametric model components with unknown regression coefficients $\boldsymbol{\gamma}$. For a Poisson GAM with logarithmic link, the mean is given by

$$\mu_i = E_i \exp\left( f_1(x_{1i}) + ... + f_J(x_{Ji}) + \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}(x_{ij}, x_{ik}) + \mathbf{z_i^t}\boldsymbol{\gamma} \right).$$

# 3 Model comparison

In this paper we consider both GLM and GAM specifications for modeling the number of deaths in a life insurance portfolio and compare the results. Model comparison using standard criteria like Akaike's information criterium (AIC) (see Burnham and Anderson (1989)) or the likelihood ratio test (see Rao (1999)) however are not advisable, since we are dealing with non nested models here. The AIC is most effective for model comparison of nested models. Brian Ripley notes "Differences in AIC are much more precisely estimated for a pair of nested models than for some non-nested pairs." See also his notes in Ripley (2004). Instead, we follow Czado, Gneiting, and Held (2009) and consider tools designed to assess model fit and predictive capabilities of a model. In particular, a non-randomized version of the probability integral transform for a discrete variable and scoring rules for count models are used.

## 3.1 Probability integral transforms for discrete random variables

After a model is fitted one is interested in adressing how sensitive the model fit is to individual observations or to groups of observations. Such an assessment can be facilitated by using a cross validation setup, where the model is refitted after deleting a single or several observations and the left out observations are then predicted using the refitted model parameters. Similarly one might be interested in assessing the prediction capabilities of a model on a set of left out observations. In both situations an observation $y$ is predicted using a predictive cumulative distribution function (cdf) $F$. The value of this predictive cdf evaluated at observation $y$ given by $F(y)$ is called the probability integral transform (PIT) (see Dawid (1984)). It is well known that if an observation $y$ arises from a continuous distribution with cdf $G(\cdot)$, then under perfect prediction, i.e. $F \equiv G$, the PIT $u := F(y)$ is an observation from a uniform distribution. For a set of observations $y_1, ..., y_K$ to be predicted using predictive cdf's $F_1(\cdot), ..., F_K(\cdot)$ the histogram of $u_k := F_k(y_k)$, $k = 1, ..., K$ is called a PIT histogram. Under perfect prediction the PIT histogram should be flat. U-shaped and bump shaped histograms indicate underdispersed and overdispersed predictions. This approach is valid if one considers continuous observations, however not for discrete observations. In the case of a discrete observation $y$, Smith (1985) considered a randomized PIT value given by

$$v := F(y-1) - u\,[F(y) - F(y-1)]\,,$$

where $u$ is an independent observation of a uniform (0,1) distribution and $F(-1) := 0$. Under perfect prediction this is an observation from the uniform distribution. The inclusion of the random quantity $u$ is not satisfactory. One contribution of Czado,

Gneiting, and Held (2009) is that they are able to construct non-randomized PIT histograms by considering the corresponding random variable

$$V := F(Y-1) - u\left[F(Y) - F(Y-1)\right],$$

and deriving the conditional cdf of $V$ given $Y = y$ as

$$H_y(v) = \begin{cases} 0 & v \leq F(y-1) \\ (v - F(y-1)) \,/\, (F(y) - F(y-1)) & F(y-1) \leq v \leq F(y) \\ 1 & v \geq F(y) \end{cases}.$$

They show for $W \sim H_y$ and under perfect prediction that $W$ has a uniform distribution. For observations $\mathbf{y} := (y_1, ..., y_K)'$ to be predicted they use the aggregated conditional cdf given by

$$H_{\mathbf{y}}(v) := \frac{1}{K} \sum_{k=1}^{K} H_{y_k}(v)$$

where $H_{y_k}$ is based on the predictive cdf $F_k$ for observation $y_k$. Finally they call the histogram of $J$ equal width bins with $j$-th bin height

$$f_j = H_{\mathbf{y}}\left(\frac{j}{J}\right) - H_{\mathbf{y}}\left(\frac{j-1}{J}\right)$$

the non-randomized PIT histogram for discrete observations. Deviations from a flat histogram indicate prediction deficiencies. For an illustration of a non-randomized PIT histogram see Figure 1 of Czado, Gneiting, and Held (2009).

## 3.2 Scores

For comparing non nested models proper scoring rules can also be used. Gneiting and Raftery (2007) consider scoring rules in order to assess the quality of a probabilistic forecast. A scoring rule $S(F, y)$ assigns a numerical value based on the predictive distribution $F$ and on the observed value $y$. We only consider strictly proper scoring rules, i.e. scoring rules for which the highest score is uniquely obtained for the true model. In the following we consider the interval and the quantile score. They are both positively oriented, i.e. the model with the highest score is to be preferred.

## 3.3 Interval scores

The interval score $IS_\alpha$ is based on a $(1-\alpha)100\%$ prediction interval $I = [l, u]$ using prediction cdf $F$. In particular, the interval score for observation $y$ to be predicted

is defined by

$$IS_\alpha(l,u,y) := \begin{cases} -2\alpha(u-l) - 4(l-y) & \text{if} \quad y < l \\ -2\alpha(u-l) & \text{if} \quad l \leq y \leq u \\ -2\alpha(u-l) - 4(y-u) & \text{if} \quad y > u \end{cases}. \tag{3.1}$$

The above definition shows that the interval score rewards narrow prediction intervals and penalizes observations which are not within the prediction interval.

## 3.4 Quantile scores

For data with many zero observations, the quantile score might be more appropriate than the interval score. The quantile score proposed by Gneiting and Raftery is based on the $\alpha$-quantile $r_\alpha$ of the predictive distribution $F$ for observation $y$. The quantile score $QS_\alpha$ is defined by

$$QS_\alpha(r_\alpha, y) := (y - r_\alpha)\left[1_{y \leq r_\alpha} - \alpha\right]. \tag{3.2}$$

For observation $y_k$, $k = 1, ..., K$ to be predicted we use the mean score given by

$$S_K := \frac{1}{K}\sum_{k=1}^{K} S\left(F_k, y_k\right) \tag{3.3}$$

where $F_k$ is the predictive cdf for $y_k$ and $S(\cdot,\cdot)$ is either an interval or quantile score function. The model with highest mean score has the highest predictive capability.

# 4 Application

For the application a portfolio of endowment life insurances with the number of deaths as response variable is examined. The data contains a number of categorical covariates including gender (**sex**, (male/female)), time since policy inception (**dur**, (0,1,...,9,10+) in years), an indicator whether medical underwriting has been conducted or not (**uw**, (yes/no)) and amount insured given in bands (**am**, (0-1,1-5,5-10,10-20,>20) in 10000 Euro). The only continuous variable given in the data is the age (18-84) of the insured.
Only a random sample of 75% of the data is used to fit the regression models, the remaining 25% is later used to assess the predictive quality of the models.
The number of deaths is modeled using both a GLM and a GAM approach. For the GLM specification, the functional form of age is investigated in advance using local smoothing methods. This non parametric function of age is then included as a covariate in a GLM. In contrast the second approach models age non parametrically in a GAM, i.e. simultaneously with the categorical variables. Model comparison and

calibration is then assessed using non-randomized PIT histograms and scoring rules as discussed in the previous section.

## 4.1 Generalized linear model

We first consider a Poisson GLM with a logarithmic link to model the number of deaths. Here GLM specifications with and without interaction effects are studied. Mortality rates typically depend on age in a nonlinear way. In particular for ages below 35 a hump, the so-called accident hump, in the rates is observed. Since age is one of the main risk drivers for mortality, it is essential for life insurance companies to model the dependency of mortality on age adequately. Within a GLM, age of the insured might be modeled by a polynomial. However, typically a polynomial of degree 6 or higher is needed to reflect the dependency on age precisely. Polynomials of higher order tend to fluctuate highly in the tail regions, an effect insurance companies would like to avoid. We therefore follow a non parametric approach to model age. While all other covariates are neglected, we first fit a Poisson GLM with age as the only covariate using local smoothing methods suggested by Loader (1999). In particular, we consider the following model

$$Y_i \sim Poi\left(E_i \, exp(f_L(age_i))\right).$$

Here $Y_i$ and $E_i$ denote the number of deaths and the exposure for age $i$, respectively. The smooth function $f_L(age_i)$ is not further specified and is to be estimated using the local smoothing algorithm implemented in the R package "locfit" by Loader (1999). The estimated functional form $\hat{f}_L(age_i)$ is given in Figure 1 together with the logarithmic crude deaths rates which are defined as

$$log(\mu_i^{crude}) := log\left(\frac{y_i}{E_i}\right).$$

Here $y_i$ denotes the observed number of deaths for insureds with age $i$.

In a second step, a Poisson GLM including the estimated function of age $\hat{f}_L(age_i)$ as covariate as well as all remaining covariates is fitted. All remaining covariates are categorical and thus are included as indicator variables. Only covariates which are significant on a 10% level are kept in the model. This might include grouping of covariate levels. In particular, the variable **dur** is grouped from initially 11 levels into only two levels. The model without interactions thus contains the covariates

$$\hat{f}_L(age),\ \mathbf{dur}\,(0-9, 10+),\ \mathbf{sex},\ \mathbf{uw} \tag{4.1}$$

In a second step, interactions are included to the model, for details see Pfettner (2009). Several interactions are found to significantly influence mortality. The re-

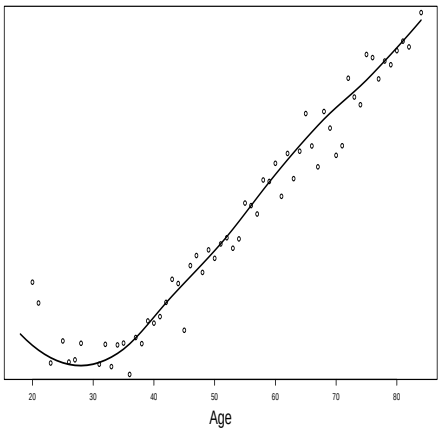


***Figure 1:*** *Estimated function $\hat{f}_L(age_i)$ (solid line) and crude logarithmic rates $log(\mu_i^{crude})$ (points).*

sulting GLM with interactions includes

$$\begin{aligned} &\text{main effects:} \quad \hat{f}_L(age), \mathbf{dur}(0-1, 2-7, 8-9, 10+), \mathbf{sex}, \mathbf{uw}, \mathbf{am}(\leq 10000, > 10000) \\ &\text{interactions:} \quad \hat{f}_L(age) \times \mathbf{dur},\ \hat{f}_L(age)\ \times \mathbf{sex},\ \hat{f}_L(age) \times \mathbf{am}, \mathbf{sex} \times \mathbf{am}, \mathbf{sex} \times \mathbf{uw} \end{aligned} \tag{4.2}$$

The GLM specificiations with and without interactions are non nested since the variable **dur** is used with different levels. For the GLM specifications utilized we aggregated the data according to the specified covariate combination levels. The aggregation is necessary to assess model fit appropriately. Table 1 displays the deviance and the degrees of freedom for the two GLM specifications before and after aggregation of the data. Without aggregation, the deviance is very small compared to the degrees of freedom which suggests the existence of underdispersion. However, after aggregating the data set, the difference between deviance and degrees of freedom is much lower. This is due to the fact, that aggregation leads to higher claim counts $y_i$ and thus more information on individual risk groups in the data.

| **Model** | **Data** | **Deviance** | **df** |
|---|---|---|---|
| GLM without interactions | not aggregated | 1533.9 | 4124 |
| | aggregated | 448.4 | 359 |
| GLM with interactions | not aggregated | 1485.3 | 4114 |
| | aggregated | 860.5 | 1029 |

***Table 1:*** *Deviance and degrees of freedom for the GLM specifications (4.1) and (4.2) before and after data aggregation.*

## 4.2 Generalized additive model

In the above approach, the data is used twice. First, when estimating the baseline mortality only depending on age, second when fitting the GLM including the estimated function of age and categorical covariates. Uncertainty in the estimation of the baseline mortality is not taken into account in the GLM specifications (4.1) and (4.2).
Generalized additive models avoid this problem by estimating non parametric functions of continuous covariates simultaneously together with the parametric components. This also allows more flexibility for modeling interactions between continuous and categorical covariates. For each level of a categorical covariate, a different functional form for the interaction with a continuous covariate may be modeled. The mean specification of the resulting GAM includes several interactions:

$$\begin{aligned} &\text{main effects}: \quad f(age), \mathbf{dur}(0-1, 2-7, 8-9, 10+), \mathbf{sex}, \mathbf{uw} \\ &\text{interactions}: \quad f_{dur}(age) \times \mathbf{dur}, f_{sex}(age) \times \mathbf{sex}, f_{am}(age) \times \mathbf{am}, \mathbf{sex} \times \mathbf{uw} \end{aligned} \tag{4.3}$$

Both the GLM with interactions and the GAM include 32 different risk profiles, i.e. 32 possible covariate combinations except age. Figure 2 displays the logarithm of the estimated mortality rates for two selected risk profiles resulting for the three considered models.
Additionally, the logarithmic mortality rates of the official table DAV2008T provided by the German actuarial association is plotted as a solid line and denoted as DAV. Note that the DAV mortality rates are the same for both risk profiles, since both risk profiles are for women and the DAV table only distinguishes between males and females but no other risk factors. The mortality rates resulting from the GLM specifications and the GAM in contrast differ significantly for the two risk profiles. For example, mortality rates in the right panel corresponding to women without medical underwriting, low sums insured and more than 10 years since policy selection are considerably higher than the rates in the left panel, corresponding to women with medical underwriting, high sums insured and less than 2 years since policy inception. Further note, that the estimated rates based on the GAM specification (4.3) and the GLM specification (4.2) including interactions are fairly close. The mortality rates estimated using a GLM without interactions in contrast differ significantly from the models with interactions for the risk profile given in the left panel. Since this model does not include interactions, only parallel shifts of the logarithmic mortality rates are possible. The models with interactions are much more flexible here. The estimated rates for all remaining risk profiles can be found in Pfettner (2009).
In order to assess which of the models performs best when considering their pre-

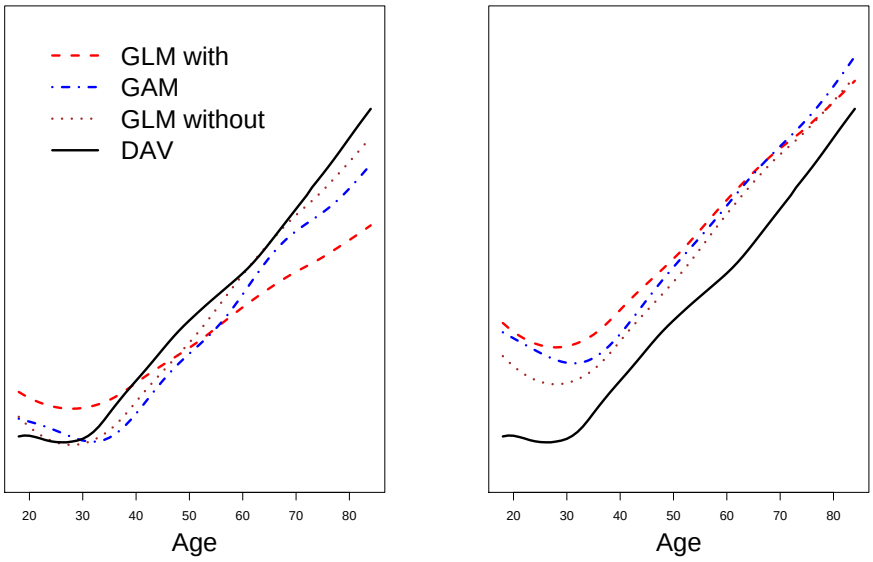


***Figure 2:*** *Fitted logartihmic mortality rates per unit exposure* $\ln(\frac{\hat{\mu}_i}{E_i})$ *as age varies for two risk profiles (left: women with duration 0-1, underwriting and amount insured* $>$ *10000, right: women with duration 10+, no underwriting and amount insured* $\leq$ *10000).*

dictive power, the non-randomized probability integral transform and scoring rules introduced in Section 3 are now considered.

## 4.3 Model assessment and comparison using PIT and scores

Probability integral transform and scores are tools to assess calibration and to compare non nested models. Generally, there are two approaches, an external and an internal one, for model assessment. We present both possibilities in this paper. The external approach corresponds to studying the predictive capabilities for "external data", i. e. new data which has not been used so far. In contrast, internal refers to the assessment of calibration where data is predicted in a cross validation setup.
For the non-randomized PIT we follow both approaches. The internal approach uses cross validation to study the goodness of fit over all data used for fitting. The external PIT in contrast uses test data and examines the predictive distribution for the test data. If the non-randomized PIT histogram is flat, then there is no evidence against the prediction ability of the model.
For scores we follow the external approach using test data only. We compare models based on the overall mean score and choose the model with the smallest mean score as best model. Model calibration can be assessed by examining the proportion of

outliers, i.e. observations lying outside the calculated prediction interval or above the prediction quantile, respectively. For a 95% prediction interval or a 95% prediction quantile, the percentage of outliers is supposed to be around 5% if the model is calibrated well.

### 4.3.1 Non-randomized PIT histograms

As mentioned above we take two approaches - an internal and an external one - to assess model fit of each of the three models. Calculation of the external non-randomized PIT is straightforward. For the calculation of the PIT, the predictive cdf for each observation in the test data is needed. For each observation $i$ in the test data the predicted mean $\hat{\mu}_i = E_i \, exp\left\{\hat{\boldsymbol{\beta}}^T \mathbf{x}_i\right\}$ is calculated, where $\hat{\boldsymbol{\beta}}$ denotes the estimated parameter vector of the corresponding model based on the selected 75 % of data. For each observation $y_i$ of the test data we use the cdf of the Poisson distribution with mean $\hat{\mu}_i$ evaluated at $y_i$ as predictive cdf for the construction of the non-radomized PIT histogram.
In order to compute the internal PIT histogram for each of the three model specifications, cross validation is used. Let $\hat{\boldsymbol{\beta}}^{[-i]}$ denote the estimated parameter vector based on a model which has been estimated without observation $i$. The predictive mean for observation $i$ is then given by $\hat{\mu}_i^{[-i]} := E_i \, exp\left\{(\hat{\boldsymbol{\beta}}^{[-i]})^T \mathbf{x}_i\right\}$. For the predictive cdf for the left out observation $i$, we use then the Poisson cdf with mean $\hat{\mu}_i^{[-i]}$. For the GAM specification, the estimated values $\hat{\mu}_i^{[-i]}$ are obtained using the "predict.gam" function of the R package "mgcv" from Simon Wood. A direct calculation of $\hat{\mu}_i^{[-i]}$ as described above is not possible, since the estimated non parametric functions cannot be assessed directly.
For the two GLM specifications the "leave one out" approach is computationally feasible, the computational costs for the GAM specification however are more than 20 times as high as for the GLM. Therefore, we decided to leave out 20 observations at a time instead of just one. Pfettner (2009) justifies this simplification by showing in Figure 25 on page 94 that the PIT does not change significantly when leaving out 20 instead of just one observation. The observations to be left out are chosen randomly and this procedure is repeated until each observation has been included exactly once in a set of left out observations.
To be precise, the estimation of the functional form of age in the GLM should have also been performed using cross-validation. Since no significant change in the estimation of the function of age is to be expected when only 20 observations are left out and to avoid the computational effort implied by a reestimation, we use the same functional form for age for each "leave-20-out" calculation for simplification. Only the GLM is reestimated in every step. The non-randomized PIT histograms for the three model specifications are displayed in Figure 3 for the external and

internal setup.
The internal PIT histograms for the models with interactions look rather uniformly distributed and therefore do not indicate any model deficiencies. The internal PIT histogram for the GLM without interactions in contrast, is slightly U-shaped and therefore suggests an underdispersed model for this data. According to these plots, both the GLM with interactions and the GAM seem to be superior to the GLM without interactions. A significant difference between the two models including interactions is not apparent.
The external PIT histograms show similar results. Here, the external PIT for the GLM with interactions (top right panel) shows the most uniform pattern, followed closely by the GAM. Again, both the GLM with interactions and the GAM seem to have a higher predictive quality than the GLM without interactions, for which the PIT is more skewed.

### 4.3.2 Scores

To complement the PIT results regarding model comparison and calibration, scores are calculated for the test data. We only consider interval and quantile scores. In order to compute the empirical quantiles of the predictive distribution needed for these scores, the predictive distribution is approximated using simulation. Prediction is done for the 25% of the data which has been neglected for model fitting.
For each observation $i$ of the test data, the predicted number of claims is calculated by $\hat{\mu}_i = E_i \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ where $E_i$ and $\mathbf{x}_i$ are taken from the test data and $\hat{\boldsymbol{\beta}}$ denotes the estimated coefficient vector of the GLM or GAM using the learning data, respectively. According to the delta method, the standard deviation of $\hat{\mu}_i$ can be estimated by $\hat{\sigma}_i = \hat{\mu}_i\sqrt{\mathbf{x}_i'\hat{\Sigma}(\hat{\boldsymbol{\beta}})\mathbf{x}_i}$ where $\hat{\Sigma}(\hat{\boldsymbol{\beta}})$ denotes the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.
The simulation proceeds as follows. For each observation $i$ to be predicted $R = 15000$ mean values $\mu_i^{*r}, r = 1, .., R$ are simulated from a Normal distribution, in particular $\mu_i^{*r} \sim N(\hat{\mu}_i, \hat{\sigma}_i^2)$. For each simulation $r$ and observation $i$ a number of claims $y_i^{*r}$ is then simulated from a Poisson distribution with mean $\mu_i^{*r}$, i.e. $y_i^{*r} \sim Poi(\mu_i^{*r})$. Prediction quantiles for observation $i$ can then be calculated as the empirical quantiles of $y_i^{*r}, r = 1, .., R$. Following this simulation approach, both uncertainty of the estimated regression parameters as well as variability of the Poisson data are taken into account. The empirical quantiles are now used to construct the corresponding mean quantile and interval scores according to (3.1) and (3.2), respectively. The overall mean score, denoted by "Mean IS" and "Mean QS" in Table 2, are calculated according to (3.3) for the test data.
For the GAM specification, again the values of $\hat{\mu}_i$ and $\hat{\sigma}_i$ are obtained using the "predict.gam" function. The resulting scores are given in Table 2.
According to the mean value of both the quantile and the interval scores, the predici-

| Mean QS | | | Outliers in % | | |
|---|---|---|---|---|---|
| GLM with | GAM | GLM without | GLM with | GAM | GLM without |
| **-0.280** | **-0.259** | **-0.834** | **4.3%** | **4.3%** | **8.0%** |

| Mean IS | | | Outliers in % | | |
|---|---|---|---|---|---|
| GLM with | GAM | GLM without | GLM with | GAM | GLM without |
| **-0.864** | **-0.481** | **-5.916** | **3.1%** | **3.7%** | **10.0%** |

***Table 2:*** *Mean values of the 95% quantile and 95% interval scores for the GLMs with and without interactions and the GAM. Additionally the percentages of outliers is given.*

tive quality of the GLM with interactions and the GAM is rather close. The percentage of outliers is also similar for both models and reasonable close to the expected 5% of outliers. For the GLM without interactions in contrast the scores are much higher and more outliers than expected are observed. This indicates that the predictive power of this model is inferior to the models with interactions and that the model is not very well calibrated yet. The scores for the GAM specification are slightly higher than for the GLM specification with interactions and therefore suggest that the GAM fit is slightly better. However, taking into account that the effort of model fitting and the computational costs are much higher for a GAM compared to a GLM, the slightly worse predictive quality might be acceptable.

# 5 Summary and conclusion

We have considered both GLMs and GAMs for modelling the number of deaths for life insurance data in this paper. In order to appropriately model the functional dependency of mortality on age of the insureds, a non parametric modelling approach was followed. In a GLM setting this involves using the data twice. In a GAM in contrast, non parametric effects are estimated simultaneously with parametric components. This also allows the modelling of different functional dependencies for interactions of age with categorical variables and thus gives more flexiblity for modeling. In this paper we examined whether the double use of data and the simplier structure in a GLM also effects the predictive ability of the model. In particular, the predictive quality of a GLM and a GAM was compared using non-randomized PIT histograms and proper scoring rules. While these tools clearly detected a GLM model without interactions as inferior to a GLM or GAM including interaction effects, model fit and calibration of the GLM and GAM with interactions turned out

to be rather close. Although prediction based on the GAM specification was slightly better than for the GLM with interactions according to the scoring rules, the higher computational costs required for fitting a GAM have also to be taken into account. Further modelling of a GAM is not as straightforward as in a GLM setting. Based on these results, the use of GLMs including a non parametric function of age which has been estimated in advance seems to be feasible.

**Acknowledgements**
We would like to thank Prof. Michel Denuit, Université catholique de Louvain, for fruitful discussions and valuable input on the use of non parametric functions within a GLM framework.

# References

Burnham, K. P. and D. R. Anderson (1989). *Model Selection and Inference: A Practical Information-Theoretic Approach 2nd Edition*. Springer.

Czado, C., T. Gneiting, and L. Held (2009). Predictive Model Assessment for Count Data. *to appear in Biometrics*.

Dawid, A. P. (1984). Statistical Theorie: The prequential approach. *Journal of the Royal Statistical Society Series A, General 147*, 278–292.

de Jong, P. and G. Z. Heller (2008). *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press.

Gneiting, T. and A. E. Raftery (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association 102*, 359–378.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. London: Chapman and Hall.

Loader, C. (1999). *Local Regression and Likelihood*. New York: Springer-Verlag.

Nelder, J. A. and P. McCullagh (1989). *Generalized Linear Models 2nd Edition*. Chapman and Hall.

Pfettner, J. (2009). Statistical Inference and Model Selection for death rate models in life insurance. Diploma thesis, Centre of Mathematical Sciences, Technische Universität München, Garching near Munich.

Rao, C. R. (1999). *Selected Papers of C. R. Rao 4th Edition*. Chapman and Hall.

Ripley, B. D. (2004). Selecting Amongst Large Classes of Models. *http://stats.ox.ac.uk/ ripley*.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting 4*, 283–291.

Wood, S. N. (2006). *Generalized Additive Models*. Boca Raton: Taylor and Francis Group.

*Claudia Czado*
*Technische Universität München*
*Zentrum Mathematik*
*Email: cczado@ma.tum.de*

*Susanne Gschlößl*
*Munich RE, Munich*
*Email: sgschloessl@munichre.com*

*Julia Pfettner*
*Email: julia@pfettner.de*

*Frank Schiller*
*Munich RE, Munich*
*Email: fschiller@munichre.com*

The opinions expressed in this paper are our own and do not necessarily reflect those of our employers.

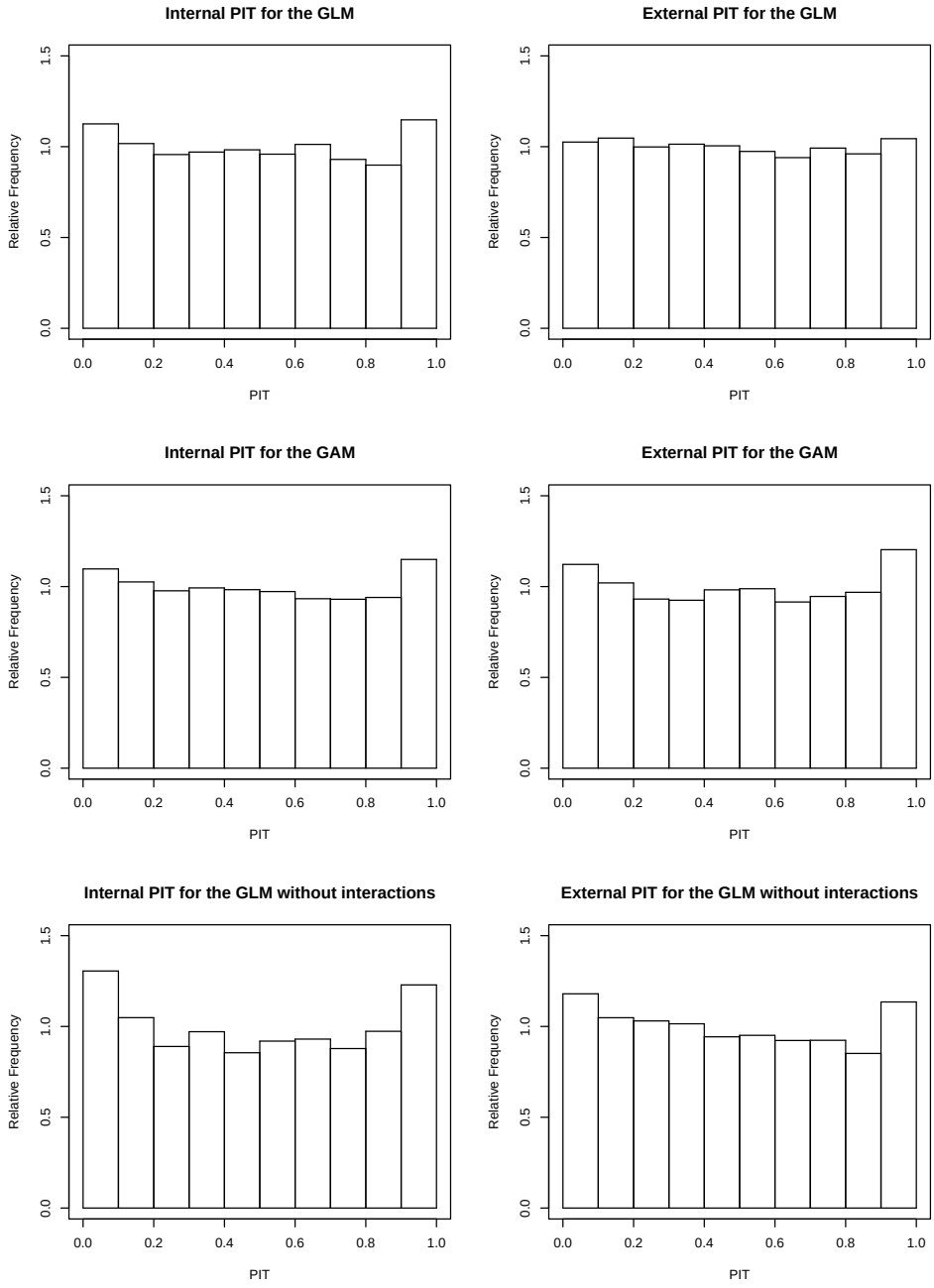


***Figure 3:*** *Comparison of the internal and external PIT histograms for the generalized linear and the generalized additive model. For the internal PIT cross validation was performed by leaving out 20 observations a time. For the external PIT no cross validation is needed since the calculations are based on the test data.*