GAM AND ROBUST EXTENSIONS IN INSURANCE RATEMAKING

ROBUST GENERALIZED ADDITIVE MODELS FOR TELEMATICS-BASED AUTO INSURANCE RATEMAKING

By STEVEN ZENG, BS

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science in Statistics

McMaster University

MASTER OF SCIENCE IN STATISTICS (2025)

Hamilton, Ontario, Canada (Mathematics and Statistics)

TITLE:	Robust Generalized Additive Models for Telematics-Based Auto Insurance Ratemaking
AUTHOR:	Steven Zeng BS (Actuarial and Financial Mathematics), McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Anas Abdallah
NUMBER OF PAGES:	xv, 70

Lay Abstract

This thesis contributes to the automobile insurance field through applications of nonparametric statistical models with data collected from a GPS device installed on the vehicle. In the introductory chapter, the steps of determining prices paid by insurance policyholders to the insurance company are outlined and discussed. The second chapter includes the theoretical background for the nonparametric statistical model and its application in the insurance industry. Chapter 3 introduces an extension over the nonparametric statistical model where data points that are seen as abnormal compared to the rest are automatically suppressed for their influence to model fit. The final chapter illustrates our contributions of applying the nonlinear statistical model and its extensions to synthetic telematics data where an improvement in model fit can be seen. Thus, insurance companies are able to come up with a more suitable price for each policyholder which would attract more customers and increase value of the company.

Abstract

In this thesis, we introduce a different P&C ratemaking strategy using telematics where complexity of telematics data is often seen as a challenge for traditional Generalized Linear Modeling. Generalized Additive Model with its flexible model structure is outlined and recent applications in the insurance industry are discussed and analyzed. A robust version of the Generalized Additive Model is then discussed where the modified penalized likelihood is able to reduce the influence of outliers present in the data. With an application on a synthetic dataset, it is shown that our results coincide with the referenced paper of Boucher (2017) and our model with the added telematics variable shows significant improvements. When outliers are introduced to the dataset, non-robust models quickly deteriorate and thus produce a poor fit whereas robust counterparts are able to maintain a similar level of model accuracy and as a result extreme risks are better identified from such policyholders. Actuaries can now utilize the added benefit of robust Generalized Additive Model for better risk classification such that a more fair pricing scheme is made possible.

To the present.

Acknowledgements

I would like to thank my supervisor Dr. Anas Abdallah for his mentoring and guidance throughout my academic career. The valuable advice he provides inspires me to pursue a career in actuarial science.

A special thank-you to Dr. Ben Bolker for answering my questions regarding outliers simulation and loss functions.

I would like to thank Dr. Jean-Philippe Boucher for helping me locate the synthetic dataset.

I would also like to thank my families for their unconditional support, love, and trust, particularly my girlfriend, my parents and my grandparents.

Table of Contents

List of Figures

List of Tables

Notation, Definitions, and Abbreviations

Notation

$g''(x)$	Second derivative of function $g()$ with respect to x

Definitions

Outliers	Data points that are seen far away from the average behavior of a point belonging to a same group

Abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
GAM	Generalized Additive Model
GCV	Generalized Cross Validation

GLM Generalized Linear Model

REML Restricted Maximum Likelihood

Chapter 1

Introduction

Insurance ratemaking is the process of determination of amount of money to be paid to the insurance company by the insured. The amount of money paid by the insured, often referred to as the premium, is calculated from the future loss of the insured estimated by the insurance company. Unlike other industries where goods or services is delivered at the moment of the payment, insurance premiums are paid some time before the service is provided. This makes calculation of insurance premiums a challenging task as the uncertainty greatly resides in the unforeseen future.

This thesis aims to provide applications in automobile insurance ratemaking in Property and Casualty (P&C), or non-life insurance which obligates the insurer to compensate the policyholder in the event of a loss incurred in relation to automobiles. Compared to the other major branch of insurance, Life Insurance, P&C exhibits a more dynamic nature where policies with less than a year of coverage appear more frequently and variables used in the pricing structure are more diverse, normally including the characteristics of the insured who is the driver of the vehicle, the driving habits of the insured, policy duration, etc., in the case of car insurance.

The industry standard method of determining P&C insurance premiums is to use the Generalized Linear Model (GLM) for its good predictive power and ease of interpretation. GLM is a statistical modeling method introduced by Nelder and Wedderburn (Nelder and Wedderburn, 1972); it is an extension of the Simple Linear Model (SLM), which only allows a normally distributed response. GLM extends the distribution of the dependent variable to follow the exponential family of distributions which conveniently covers many more distributions such as the discrete distributions of Poisson distribution, Binomial distribution and continuous distributions such as Gamma distribution. Generalized Linear Models for Insurance Rating 2^{nd} (Goldburd et al., 2016) showcases the steps of building a GLM model for insurance rate estimation in detail. This monograph also describes model building in general by going through data preparation, model fitting and adjustment, model evaluation and validation.

One approach frequently used in practice to estimate P&C insurance premiums is to view the pure premium (i.e., the premium before adding profit margin and overhead costs) as the product of two components: the frequency of the claim and the severity of the claim. The frequency of the claim describes how many times the insured incurs a loss; it is a nonnegative integer. The severity of the claim is a monetary amount depicting the seriousness of an average claim made by the insured. In order to estimate pure premium using aforementioned strategy, two models are fit separately on claim frequency and claim severity and then the estimated pure premium is obtained by multiplying the frequency estimate with the severity estimate. The estimates of both claim frequency and claim severity can be performed using GLM in software like R (R Core Team, 2023), SAS, and Python.

Compared with claim severity, which is often found to be difficult to estimate because

of its uncertainties associated with catastrophic events and multi-period claim settlement process, claim frequency appears easier to work with. Claim frequency is a special representation of claim count; it is the claim count divided by exposure which is normally taken as coverage duration for insurance policy. To model claim count using GLM, we often set the family of distribution to Poisson with log link structure to enable a multiplicative effect between the response and the covariates. In a similar fashion, claim frequency is modeled where claim count acts as the response but exposure is treated as an offset.

Exposure, along with characteristics of the insured such as age, gender and marital status and characteristics of the car such as the age of the car, the model of the car and the fuel type of the car, constitutes the traditional rating variables in auto insurance rating structure. With the help of technology and digitization, more and more advanced tools are being invented and implemented in the insurance industry. Telematics is a prominent example of this which acts as a bridge between the actuary and data that are challenging or even impossible to collect in a traditional rating procedure.

If they agree to install a GPS device on their vehicle, policyholders are often offered a discount in insurance premiums with the exchange of driving behavior data sent to the insurance company on the fly. This gadget collects driving patterns and anomalies and sends the data to the insurer. The granularity across driving behavior data collected in the form of telematics makes more advanced fair pricing structure and risk segmentation possible. Such data is collected in the forms of driving activity variables such as distance driven, brake and acceleration intensities, and whether the insured is driving in rush hours. The added information obtained from this device

requires careful treatment; a more advanced modeling process is needed to gain the extra insights from these data.

The Generalized Additive Model or GAM adds this needed flexibility over GLM, at the same time maintaining a level of interpretability over machine learning methods such as Neural Networks and Random Forest. Thus, this makes GAM a potential candidate for nontraditional insurance data where telematics information is available. Similar to GLM, nonlinear relationships between the response and the covariates are allowed except that the covariates are now represented by nonparametric smooth terms. Therefore, with the added flexibility GAM enables data to speak for itself. An overview of GAM and a literature review of GAM in insurance are provided in the next chapter (Chapter 2) whereas an application of GAM with automobile count data is illustrated in Chapter 4.

Outliers are almost always present in real-life data; actuaries are often tasked with predictive modeling that involves this type of data where the presence of outliers would jeopardize model accuracy. In an attempt to deal with this challenge, we introduce a robust extension of GAM which in theory is more resilient to deviations from model assumptions caused by outliers and hence would produce a better fit. It behaves like a filter where it automatically selects and down-weighs the influence of points that are seen as abnormal from the rest of the data. It is a robust method in a sense that ideally it will perform similarly to non-robust models with no outliers and will outperform non-robust models with outliers present. Chapter 3 will provide theoretical background on robust GAM and its applications in the insurance industry. In Chapter 4, an application of robust GAM with the same telematics data will be presented as an extension to traditional GAM in the framework of claim count

prediction followed by Chapter 5 which concludes the thesis.

Chapter 2

GAM Models

2.1 Review of Linear Model and Generalized Linear Model

The traditional Simple Linear Model (SLM) is a regression method that is used to find linear relationships between the response and a single covariate. It has the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where y_i represents the response or dependent variable of interest and x_i is the single covariate. It also assumes that the random error term ϵ_i is independently and identically distributed as a Normal distribution with mean 0 and variance σ^2. The β_0 and β_1 are model coefficients they can be estimated by Least Squares (LS) method where

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

is minimized. The LS estimates of β_0 and β_1 are

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\widehat{\beta_0} = \bar{y} - \beta_1 \bar{x}$$

where the estimates are found by taking the derivatives of the objective function with respect to β_0 and β_1 and then setting the results to 0.

An extension of this is the Multiple Linear Model (MLM) where the response is modeled by a group of predictor variables which expressed in matrix form as

$$\vec{Y} = \vec{X}\vec{\beta} + \vec{\epsilon}$$

where again $\vec{\epsilon}$ follows as Normal distribution with mean 0 and a variance covaraince structure of Σ with diagonal values being σ^2. As in the case of SLM, the coefficient estimates can be found by Least Squares

$$\widehat{\vec{\beta}} = (X^\top X)^{-1} X^\top Y$$

The Generalized Linear Model (GLM) introduced by Nelder and Wedderburn (Nelder and Wedderburn, 1972) extends MLM by enabling $\vec{Y}|\vec{X}$ to follow a distribution from Exponential distribution family. With GLMs, we are no longer limited to Normal distribution and the various choices of distributions such as Poisson distribution, Binomial distribution and Gamma distribution create a powerful toolbox for modeling of data across different scenarios. The coefficient estimates are found through

maximization of likelihood and it is usually done using an iterative method such as Newton-Raphson or Fisher Scoring.

2.2 Introduction to Generalized Additive Model

One might pose a question like this: is there a modeling method that can save us the efforts of specifying a particular model form? Generalized Additive Model (GAM) is an answer to this. Introduced by Hastie and Tibshirani in 1986, GAM enables us to estimate the relationships between the response and the predictor variables in a nonparametric manner while maintaining a likelihood based modeling procedure and making normal statistical inference possible. In the case of a single predictor

$$y_i = s(x_i) + \epsilon_i$$

where $s(\cdot)$ represents a nonparametric smooth function that is allowed to fit automatically by the data. That is, without the need to specify a parametric form of the curve, we are able to fine tune the smoothed curve to capture the local patterns of the data.

This is done by introducing the concept of neighborhood w. w which is defined as the span of the model describes the size of group of points that are recognized to be close to each data point and is used to estimate the local effects of the predictor. With w specified, covariate values will be sorted in ascending order where half of $\lfloor nw \rfloor$ (n, the number of total data points) will be included to the left of x_i and the other half to the right symmetrically in the estimation process. Span usually takes value between 0 and 2 with 0 only including the point itself resulting in a perfect fit to the data and

2 enabling the inclusion of all data points for every neighborhood which is in fact the least squares fit under the running lines scatterplot smoother (i.e., $\widehat{s}(x_i) = \widehat{\beta_{0i}} + \widehat{\beta_{1i}} x_i$ fitted locally to neighborhood of point x_i). Therefore, the span specifies the size of the neighborhood of each point where scatterplot smoother comes up with the local estimate of the covariate effects to the response.

The determination of the span w is an automatic process where it is chosen to minimize Cross Validated Sum of Squares (CVSS)

$$CVSS(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{s}^{-i}(x_i))^2$$

where $\widehat{s}^{-i}(x_i)$ is the smooth estimate of point x_i without including it during the fitting process.

One example of the scatterplot smoother is the local average smoother or running average smoother where arithmetic mean is calculated over the span which means $\widehat{s}(x_i)$ is estimated to be the average value of corresponding y_i around x_i. However, due to the weird behavior of it around the end points, running lines smoother is often preferred (Hastie and Tibshirani, 1986).

The usual Fisher Scoring fitting algorithm in GLM is used with modification that the updating formula is estimated locally using the smoother. The iterative process still proceeds as before until convergence. However, here we use Local Scoring or Local Likelihood procedure where the solution to the score equation still follows from the Taylor series expansion but now estimated locally by the smoother (Hastie and Tibshirani, 1986).

Moving into the multiple covariate case, instead of estimating a single smoothed function for all of the predictors which would be problematic because of the curse of

dimensionality where neighborhood breaks down and points become sparse in the case of high dimensions, we estimate smoothed functions for every covariate one by one sequentially. This is called the Backfitting algorithm where we iterate through the process of estimating smoothed function effect for $s_j(x_j)$ ($j = 1, ..., p$ covariate) using partial residuals in the case of Gaussian distribution and using adjusted dependent variable in other cases (Hastie and Tibshirani, 1986).

2.3 Modern Development of GAM

Despite the appeal of the Backfitting algorithm in allowing different smoothers to be selected for each smoothed terms, it becomes difficult to incorporate methods for smoothness control such as Generalized Cross Validation (GCV).

This section introduces the concept of building GAM using bases which is described in detail in the book by Wood (2017) and the accompanying R package `mgcv`.

As in the previous section, the general model form of GAM outlined by Simon Wood is

$$g(\mu_i) = \vec{X}\vec{\beta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i})$$

where $g(\cdot)$ is the link function, $\vec{X}\vec{\beta}$ is the parametric effect term, and $f_i(\cdot)$ is the smoothed effect term. The idea is to represent the smoothed functions $f_i(\cdot)$ with basis expansions.

Let us first look at the simple univariate smooth case where a single smoothed function is required to be estimated

$$y_i = f(x_i) + \epsilon_i$$

.

To estimate smooth function $f(\cdot)$, we first need to select a basis for which a set of functions is treated as known and is believed to be a close approximation to the true $f(\cdot)$ so that

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j$$

where $b_j(x)$ is the j^{th} basis function and β_j is the corresponding coefficient. That is, we aim to transform nonparametric smooth function into linear representatives. The simple choice is the polynomial basis where an m^{th} order polynomial basis result in

$$f(x_i) = \beta_0 + \beta_1 x_i + ... + \beta_m x_i^m$$

.

Another basis, the piecewise linear basis, is often used in place of polynomial basis due to its improved performance near end points. The concept of w, the neighborhood, is defined in previous section and introduced by Hastie and Tibshirani to measure the number of points that are chosen to be close to each datum in smooth function so that the local behavior could be estimated. But here we introduce a similar concept of knots for realization of closeness. Knots represented here by x_j^* where $j = 1, ..., k$ are points where linear basis functions connect. They are the locations of function discontinuities where derivatives become discontinuous. $b_j(x)$ then will be formed with tent functions for which the functional value $b_j(x)$ of an arbitrary point x only

depends on adjacent knots x^*_{j-1} and x^*_{j+1} on either side of the point x (i.e., $x^*_{j-1} < x < x^*_{j+1}$) (Wood, 2017, Section 4.2.1). $f(x)$ then becomes the linear model

$$f(x) = \vec{X}\vec{\beta}$$

where $x_{ij} = b_j(x_i)$. As it can be seen above, now our smooth function is written in the form of parametric terms.

Besides choosing a basis to approximate the smooth function, function wiggliness also needs to be quantified and measured so that model obtains an appropriate degree of smoothness. What is normally done in practice is to decide on a basis dimension k or number of knots that is more than needed and then introduce a penalty parameter λ to control function wiggliness. Therefore, rather than minimizing LS we have

$$||Y - X\beta||^2 + \lambda\beta^\top S\beta$$

which is the objective function in penalized regression. The second term in the objective function above is used to measure function wiggliness and usually can be approximated by squared second order difference $\sum_{j=2}^{k-1} = (f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1}))^2$. We then can proceed similarly as we would in penalized regression method and obtain coefficient estimates

$$\widehat{\beta} = (X^\top X + \lambda S)^{-1} X^\top Y$$

and the influence matrix $A = X(X^\top X + \lambda S)^{-1} X^\top$.

Before computing coefficient estimates β's, we first need to decide which value λ the smoothing parameter should take. This is achieved through cross validation where

$$M = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_i - f_i)^2$$

is minimized. M here measures the average squared difference of our estimated smooth function with the true function and since the true function is unknown it is often approximated by computing OCV (Ordinary Cross Validation) score

$$OCV = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_i^{[-i]} - y_i)^2$$

where $\widehat{f}_i^{[-i]}$ is the fitted estimate of datum i without i during the fitting process. It can be shown that $E[OCV] = E[M] + \sigma^2$ which is the expected squared error of predicting new data. Generalized Cross Validation (GCV) score is often used in place of OCV for its ease of computation and invariance property

$$GCV = \frac{n\sum_{i=1}^{n}(y_i - \widehat{f}_i)^2}{(n - tr(A))^2}$$

where A is the influence matrix.

Such GAM can also be written in the forms of Bayesian model and mixed model for which marginal likelihood maximization or Restricted Maximum Likelihood (REML) could be used to estimate σ^2 and λ.

For multiple variates, the similar additive model could be constructed but now with multiple penalty terms correspond to each smoothed covariate while a constraint $\sum_{i=1}^{n} f_j(x_i) = 0$ is required for each smoothed function $f_j(\cdot)$ to ensure identifiability of each term. Penalized likelihood maximization is used for distributions other than Gaussian distribution and this is typically done in practice via Penalized Iteratively Reweighted Least Squares (PIRLS).

Next, we will introduce cubic regression spline basis and tensor product basis which have more appealing properties compared to the piecewise linear basis.

It can be shown that the cubic smoothing spline basis is the smoothest univariate smoother minimizing our score objective where it achieves the lowest squared error and the lowest smooth penalty term. Therefore, ideally we would want to build our smoother with cubic smoothing splines. But due to the computation complexity of the cubic smoothing spline where every datum is needed during the fitting process, we turn to the cubic penalized regression spline, which unlike cubic smoothing spline, only requires a set of k knots to be used for fitting. The rest of the data can be fitted accordingly based on the fit at the knots. Univariate base other than cubic penalized regression spline basis could be used in different scenarios. Cyclic cubic regression spline is suitable for data with seasonal trends and this is achieved by constraining on an equal functional value as well as first and second derivatives at the boundaries around the knots, whereas the P-spline basis is constructed by putting smoothness penalty directly on coefficient β's and it is particularly useful in Bayesian GAM where its low rank nature provides computational efficiency.

For modeling interactions among multiple covariates, tensor product basis will be used. Tensor product basis is one popular example of multivariate smooth which allows straightforward generalization of multiple variables smoothing while maintaining scale invariant property. This is achieved by making the coefficients of each smooth change with smooths of other variables. Thus, it enables the tensor product smooth to model variables with different scales which is used later in subsection 4.2.2 where distance traveled (in km) is modeled along with policy duration (in years).

2.4 GAM in Insurance

The uses of GAM in insurance has been less popular than GLM. Particularly, the focus of GAM applications in the insurance industry has been on ratemaking where the benefit of flexible nonparametric modeling of GAM can be taken advantage of to discover complex relationships between the response and predictors that are seen either to be difficult or impossible to uncover with a traditional GLM.

In life insurance, GAM is used to incorporate nonlinear effects of covariates. In Czado et al. (2009), Poisson GLM and GAM are used to model number of deaths with smoothed effects of age and its interactions with other variables for endowment life insurance policies where it is found that GLM performs similarly to GAM. A natural cubic spline basis GAM is used in Mau et al. (2020) with Gamma distribution to model the relationships between healthcare costs and days since initiation of Chemotherapy treatment for patients diagnosed with a type of Leukemia in USA.

The implementation of GAM in P&C insurance is more prominent compared with the health insurance industry with emphasis on automobile insurance pricing. There are dozens of references pointing to the use of GAM in non-life insurance which are all written after the year of 2000.

Chen et al. (2023) implements cubic spline GAM to model dependent structure in aggregate claims with applications in simulation as well as an auto insurance dataset where it shows that both dependent GAM with either frequency and severity modeling or pure premium Tweedie modeling performs better than the GLM counterparts with a higher adjusted R^2 and a lower AIC. The comparison of differences in approaches of actuarial modeling of frequency and severity using GLM and GAM in general with a Spanish auto insurance dataset is outlined in Díaz Martínez et al. (2023). It is found

that despite the excess of computation time of GAM compared with GLM, GAM can take advantage of additional information and thus produce an improvement in predictions. The utilization of Bayesian P-spline GAM can be found in Fahrmeir et al. (2003) where it is used to model frequency and severity using Poisson and log-normal with respect to spatial covariates with application in auto insurance data in Germany. Novkaniza et al. (2025) considers cubic spline GAM with Poisson distribution to model claim frequency based on distance traveled and policy duration with application in the same auto insurance data as ours but with three knots selected and placed at 0, 0.5 and 1 quantiles for each predictor. The relative risks are derived in a tariff structure where the similar estimates and trends that the predicted risk increases for corresponding distance traveled as duration increases are observed. Roznik et al. (2019) uses thin plate spline in GAM to predict air temperature in agricultural insurance with the goal of reducing interpolating error due to sparse weather stations for better risk estimation in index based weather insurance. The model incorporates elevation data and found that GAM performs on par with Universal Kriging where both performed better than Nearest Neighbor in terms of a lower RMSE and a lower MAE with a real data set in Alberta, Canada. Cubic spline GAM and GLM are implemented in Tingting (2018) to estimate pure premium in a dependent frequency severity structure where it is shown that the dependent GAM model performs better in estimating pure premiums with both applications on simulated data and one year auto insurance claims. Xie (2024) uses GAM and GLM with territory (e.g., urban or rural) and class (i.e., types of use) variables to model aggregated loss cost and premiums for different coverage types where it is found that GAM performs better with a higher adjusted R^2. The application of GAM for pure premium estimation with addition of UBI data can

be found in Xie and Shi (2023) where the same telematics data as ours is included with territorial variable generated using K-means clustering. The incorporation of territory variable improves the original model resulting in a lower GCV score and a lower AIC. In Kaivanipour (2015), the author investigates the performance of L-curve which chooses smoothing parameter with a balance between regularization (i.e., model smoothness) and deviance (i.e., model fit) with cross validation in choosing smoothing parameter in GAM. Although L-curve excels in computation efficiency, it suffers from under smoothing compared with cross validation using non-life test data. Finally, we have Boucher et al. (2017) which is our main paper of discussion where a cubic spline basis and a tensor product basis GAM are implemented with application in telematics motor insurance data.

Other than applications of GAM in insurance, there are a few uses of GLMM (Generalized Linear Mixed Model) and GAMLSS (Generalized Additive Model for Location, Scale and Shape) in non-life insurance field where random effects in GLMM and multi-parameters modeling in GAMLSS are seen as potential improvements over GAM. Seyam and Hussien (2022) implements GLM, GLMM, and GAM to estimate pure premium for an auto insurance company in Egypt where it is found that GLMM is the most preferable as it has the lowest AIC. In Pitt et al. (2020), GAMLSS is used to model mean and scale parameters for aggregated loss using a compound Poisson distribution with a heavy tailed Beta distribution for catastrophe risks in natural perils where they find GAMLSS performs better comparing to GLM in estimating conditional distribution based on both graphical representation and Kolmogorov Smirnov test. Denuit and Lang (2004) develop and analyze a Bayesian GAM with incorporation of spatial variables for non-life ratemaking using a Belgian motor insurance data.

Klein et al. (2014) modifies Bayesian GAMLSS to capture effects from geo-spatial data and its interactions on insurance premiums with help of the same Belgian dataset. Distributions such as Negative Binomial and zero-inflated Poisson with random effects are used to take into account of group-specific variations for claim frequency and zero adjusted log-normal, Gamma, and inverse Gaussian are used to model zero amount claims directly for severity.

Chapter 3

Robust GAM Models

3.1 Introduction to Robust Statistics

The concept of robustness arises from dealing with deviations to model assumptions. As noted in Casella and Berger (2001), Huber explains the idea of robustness by emphasizing its three characteristics. Firstly, the robust statistical procedure should possess a good efficiency under assumed model. Secondly, small deviations from model assumptions should only damage performance slightly. Thirdly, relatively large deviations from model assumptions should not destroy model performance. In short, a robust model is one such that performs on par with non-robust models under model assumptions and is more resilient of mild to large deviations than non-robust counterparts.

One example of a robust estimator can be seen through finding an alternative estimator from sample mean and sample median. This estimator, considered by Huber

(Casella and Berger, 2001), is found by minimizing

$$\sum_{i=1}^{n} \rho(x_i - a) \tag{3.1.1}$$

where $\rho(\cdot)$ is a function characterized as

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq k \\ k(|x| - \frac{1}{2}k) & \text{if } |x| \geq k \end{cases} \tag{3.1.2}$$

with k being a tuning parameter controlling the robustness of the estimator. It can also be seen that if $\rho(\cdot)$ is the square function then 3.1.1 becomes $\sum_{i=1}^{n}(x_i - a)^2$ which is the least squares objective and the sample mean estimator achieves the minimum. If on the other hand we take $\rho(\cdot)$ as the absolute value function, then 3.1.1 becomes $\sum_{i=1}^{n} |x_i - a|$ and as a result the sample median minimizes this objective. Close examination of $\rho(\cdot)$ in 3.1.2 shows that the function is continuous and differentiable at $|x| = k$ where $\rho(x) = \frac{1}{2}k^2$.

With $\rho(\cdot)$ specified in 3.1.2, it can be observed that the objective function in 3.1.1 would behave similarly to the least squares objective with $|x| \leq k$ and behave close to the absolute value objective where the effect of outliers is diminished when $|x| \geq k$. Hence, this further justifies the fact that Huber wants it to be the middle ground between the sample mean estimator and the sample median estimator. As the value of k increases, the resulting estimator approaches the sample mean estimator.

In general, the estimator minimizing 3.1.1 is termed the M-estimator (Maximum Likelihood Estimator) for which a particular $\rho(\cdot)$ is specified. In the case of $\rho(\cdot)$ being set to the negative log likelihood, the estimator becomes the MLE. The efficiency of the robust estimator can be compared with the efficiency of the non-robust estimator

by computing ARE (Asymptotic Relative Efficiency)

$$ARE(\widehat{a_{robust}}, \widehat{a}) = \frac{\text{Asymptotic Variance of } \widehat{a}}{\text{Asymptotic Variance of } \widehat{a_{robust}}}$$

It is shown that the ARE of an M-estimator to the MLE is always less than or equal to 1 meaning that the M-estimator is always less efficient than MLE and it trades off between efficiency and robustness (Casella and Berger, 2001).

3.2 Introduction to Robust Generalized Additive Model

In Aeberhard et al. (2021), a robust procedure for GAMLSS was developed where robustness is achieved through modifying the likelihood function inside the objective function. GAMLSS was introduced by Rigby and Stasinopoulos (2005) and similarly to GAM it allows both the parametric and nonparametric effects to be modeled with the response. In addition, GAMLSS enables us to model parameters of a distribution other than the location parameter and this enables the modeling for distributions outside the exponential family of distributions. Thus, parameters such as scale and shape of a distribution can be modeled

$$g_\sigma(\sigma) = \vec{X_\sigma}\vec{\beta_\sigma}$$
$$g_\nu(\nu) = \vec{X_\nu}\vec{\beta_\nu}$$

where $g_j(\cdot)$ is the corresponding link function for j^{th} parameter and $\vec{X_j}\vec{\beta_j}$ is either the parametric effects term or parametric representation of the smoothed effects term.
It is described in Aeberhard et al. (2021) that the robust GAMLSS proposed here

applies robustness directly on the penalized likelihood objective function. And it is said to be an improvement over other robust GAMLSS procedures in the literature since the sampling distribution can be derived for inference and the direct application on the objective function enables natural derivation of robust information criterion for choosing the smoothing parameter. Normally, in penalized likelihood we would have

$$l_{penalized}(\mu, \sigma, \nu) = l(\mu, \sigma, \nu) - \frac{1}{2}\delta^\top S\delta \tag{3.2.1}$$

where δ is the parameter vector (μ, σ, ν) and S involves smoothing parameters λ_j's. Therefore, the first term on the right hand side is the normal log likelihood and the second term is the term controlling wiggleliness of the fitted line. Our goal then is to maximize 3.2.1 or minimize negative 3.2.1.

The magic of robustness happens at the level of log likelihood where normal log likelihood $l(\mu, \sigma, \nu)$ is replaced with its robust version

$$\tilde{l}(\mu, \sigma, \nu) = \sum_{i=1}^{n} \rho_c(l(\delta)_i) - b_\rho(\delta) \tag{3.2.2}$$

for which $\rho_c(\cdot)$ is specified to control contributions of data points to the likelihood value and $b_\rho(\delta)$ is a correction term for Fisher consistency. $\rho_c(\cdot)$ used here in robust GAMLSS is

$$\rho_c(z) = \log \frac{1 + \exp(c + z)}{1 + \exp(c)}, \text{where } c > 0$$

where c acts like a tuning parameter for robust fitting. It can be further observed that the derivative with respect to z, $\rho_c'(z)$, acts like a weighting function for each

data point and ranges from 0 to 1 as it is the logistic function.

A modified version of the trust region algorithm can be used to maximize 3.2.1 where it exhibits stable convergence. `gamlss()` function in R package `GJRM` is built for fitting such models (Marra and Radice, 2023). The selection of smoothing parameter λ_j can be chosen by minimizing the robust version of AIC and BIC

$$\begin{aligned} rAIC(\lambda) &= -2\tilde{l}(\delta) + 2\text{tr}[(\widehat{M(\delta)} + S)^{-1}\widehat{Q(\delta)}] \\ rBIC(\lambda) &= -2\tilde{l}(\delta) + \log(n)\text{tr}[(\widehat{M(\delta)} + S)^{-1}\widehat{Q(\delta)}] \end{aligned} \tag{3.2.3}$$

where the second term on the right hand side can be interpreted as the total effective degrees of freedom from fitting the smoothed model. $\widehat{M(\delta)}$ and $\widehat{Q(\delta)}$ are the observed values of $M(\delta) = -E[\tilde{l}''(\delta)]$ and $Q(\delta) = E[\tilde{l}'(\delta)\tilde{l}'(\delta)^\top]$. Since the smoothing parameter selection based on rIC (robust Information Criterion) described here involves iteration of a two step procedure where we first optimize λ given values of $\hat{\delta}$ and then we update parameter estimates $\hat{\delta}$ based on selected smoothing parameter, this process would become rather slow. Hence, an alternative procedure, a robust version of Fellner-Schall (Wood, 2017, section 6.4) would be used for efficient computation.

To illustrate model improvements, the authors first use a simulation dataset of two covariates constructed by randomly selecting 5% of the MRI brain imaging data studied in Wood (2017) and enlarging the corresponding response by adding 10 comparing to a response median of 0.86. A robust Gamma thin plate regression spline basis model is fitted with a non-robust alternative and the result shows that the robust model produce a much better fit with a lower MSE whereas the non-robust version produce a large positive bias. The same promising results can also be concluded through another set of simulation data where other robust GAM alternatives in the literature

are fitted. The robust GAMLSS performs on par with other robust GAM alternatives and shows a lower MSE and a lower deviation among others.

3.3 Robust GAM in Insurance

There have not been many applications of robust GAM in insurance to our best knowledge. There appears to be only one example of Chang et al. (2024) recently where a robust GAM is developed for claim reserves estimation. The robust GAM is used to model incremental claim amount with smoothed effects of accident year and development year in a loss triangle. The model uses cubic regression spline setting each accident year and development year as knots. The robustness of the model is achieved by modifying the score equation of the penalized likelihood where a winsorizing bounded function is introduced to limit the influence of extreme values in the response. Finally, the robust reserve estimates are determined by stratified bootstrapping where residuals are drawn randomly with replacement in each strata and this ensures the outliers would be sampled equally likely as the original data.
In addition, the robust GAM is fitted along with GLM, GAM, and robust GLM on a simulation data and a real dataset under both scenarios of outliers absent and outliers present. The results show that the robust GAM developed performs similarly to GAM and both perform better than GLM and robust GLM without contamination. Moreover, the robust GAM achieves both the lowest MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) among others by a large margin and thus performs the best among all four models with contamination introduced where one preset incremental claim is enlarged 10 times.

Chapter 4

Empirical Illustrations

Boucher et al. (2017) investigate the effect of distance driven and policy duration on claim count using GAMs. Unlike traditional insurance pricing strategies where policy duration is used as the only factor determining accident risks of each insured, the paper suggests the addition of annual distance driven, which is obtained through an onboard GPS device. With the help of the GPS device, more telemetics data such as annualized percentage time on the road, number of sudden accelerations and number of sudden brakes can be obtained.

First, we introduce the used synthetic automobile claims data. The preliminary descriptive analysis shows the nonlinear relationship between claim frequency and annual distance driven, whereas the relationship between claim frequency and policy duration is linear. Next, two Poisson GAMs are modeled and compared where the first models annual distance driven and policy duration separately to predict claim frequency using independent cubic splines and the second models them together by incorporating a tensor product basis. A third model is introduced as a Poisson GLM setting policy duration as an offset and transforming annual distance driven as an

ordered factor variable. The predictive results of the three aforementioned models are compared through Mean Absolute Residuals (MAR) computed using a validation dataset. The conclusions are that the GLM model obtains similar predictive power by achieving similar MAR as the other two GAMs but produces a more volatile distribution of absolute residuals. In the last section of the paper, results of two new GAMs mimicking traditional GLM pricing structure through adding knots and customizing knot locations are shown. The main contributions of the paper lies in the addition of telemetry variable annual distance driven along with traditional rating variable policy duration in the process of predicting claim frequency and the connection between old school GLMs and newly developed GAMs.

This chapter is structured as follows. An introduction to the synthetic dataset used fills section 1. Section 2 details the replication results according to flow of the original paper. Section 3 concludes the remarks from comparing and contrasting the results of ours to those of the paper. Finally, the last two sections extends the GAM model on the ideas of adding more telematics information and introducing a robust GAM method.

4.1 Data Description

The dataset being used in Boucher et al. (2017) belongs to an insurance company in Spain and is not available in public. However, luckily So et al. (2021) create a synthetic telematics dataset based on data used in Boucher et al. (2017). The synthetic dataset greatly resembles the real dataset by design and thus it is possible to re-perform the analysis and to replicate the results from the original paper.

The synthetic dataset consists of 100,000 insurance policy entries and 52 variables.

Variable	Description	Type
Duration	policy duration in days	continuous
Insured.age	age of the insured in years	continuous
Insured.sex	sex of the insured	categorical
Annual.pct.driven	annual percentage time driven	continuous
Pct.drive.mon	percentage driven on Mondays	continuous
NB_Claim	claim count of the insured	discrete
Total.miles.driven	miles driven during coverage	continuous

Table 4.1: Description of some variables in synthetic dataset.

The 52 variables include claim count variable, insured characteristics variables such as insured's age, insured's sex and insured's credit score, and telematics variables such as annual distance driven, annual percentage of driven on days (Mon/Tue/Wed etc.) of the week, annual number of turning intensities and so on. Table 4.1 gives a glimpse to characteristics of a few variables. So et al. (2021) give a detailed descriptions of each variable.

Additionally, similar to *Table 1* from the original paper, we have Table 4.2 where distributions of a few representing variables are described. According to Table 4.2, the distribution of annual distance driven is very similar to what the original data has but insured's age and vehicle's age have quite different distributions than the ones from the original data. Particularly, insured's age has an average of 51.38 years compared to the much younger age of 25.97 years of the original one. Vehicle age has an average of 5.64 years and a minimum of -2 years compared with an average of 7.91 years and a minimum of 0 years of the original. It is mentioned in the original paper that the younger age of 25.97 years is the result of a selection of policyholders who agree to have the GPS device installed in the car and thus this makes telematics information available. The average driving age of Spain (Boucher et al., 2017) is

Variable	Average	Standard Deviation	Minimum	Maximum
Total.miles.driven	4833.6	4545.943	0.1	47282.6
Insured.age	51.38	15.467	16	103
Car.age	5.64	4.06	−2	20
NB_Claim	0.04494	0.218	0	3

Table 4.2: Distribution of representing variables.

around 48.63 years which is much older than the one in the paper but very close to what we have in synthetic dataset. The negative minimum value of vehicle's age in synthetic data set is explained by purchase of newer model cars in advance. Since negative values may influence our results and all values of vehicle's age in the original paper are non-negative, a value of 2 years are added so that the synthetic dataset has a minimum vehicle's age of 0 years and a mean value of 7.64 years.

Gender composition of our data is very similar to the one from the original paper as we have 46% of females and 54% of males compared to 46.3% of females and 53.7% of males in *Table 2* of the paper. Nonetheless, parking type variable is missing completely from our dataset and it is a variable describing whether the insured parks the car outside or inside a private garage. A more complete and detailed analysis can be found in So et al. (2021).

It can be seen from above preliminary data analysis that most of the variables in the synthetic dataset resemble patterns of the variables from the original dataset but with a few exceptions. Therefore, it may be of concern that the difference in patterns of a few variables might deviate our replication results from the original paper and this will be investigated thoroughly in the next section where replication process takes place.

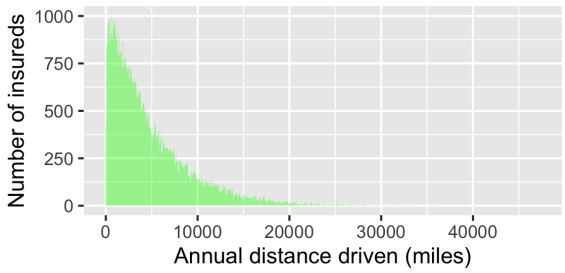

Figure 4.1: Distribution of annual distance driven (around 80 miles per bin)

4.2 Referenced Model with Synthetic Data

In order to assess the predictive power of each model, we use a train-validation set approach where stratified sampling is done to ensure that a train-validation split of 70-30. Before the replication starts, we examine the distributions of both annual distance driven and policy duration in Figure 4.1 and Figure 4.2. The distribution of annual distance driven looks almost identical to *Figure 1* in original paper where the distribution has a mode at 0 then decreases gradually as annual distance driven increases. Most of the insureds drive less than 20,000 miles during coverage, which is also in accordance with original paper. However, the distribution of policy duration greatly differs from *Figure 2* of original dataset where there is only one mode at policy duration of 1 year with the rest of data exhibiting a uniform distribution; here we have two modes at policy duration of 0.5 years and 1 year respectively in synthetic dataset. This difference in distribution of policy duration can possibly be explained by employing a U-shaped distribution instead of a uniform distribution during synthetic data generation process (So et al., 2021).

Descriptive analysis regarding distribution of claim counts arrives at the same results as the original paper. That is, the distribution of claim frequency displays an increasing but nonlinear pattern with annual distance driven (Figure 4.3). And the

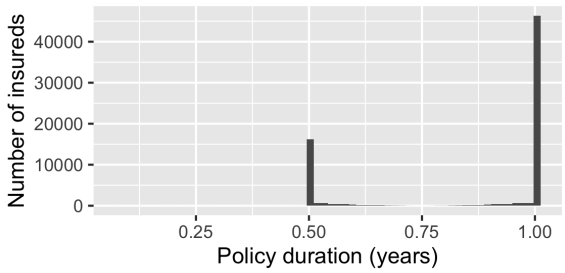

Figure 4.2: Distribution of policy duration (around 6 days per bin)

distribution of claim frequency increases linearly as policy duration increases (Figure 4.4). Although preliminary analysis gives us an idea of what the data looks like, more detailed predictive modelling is done next to investigate the effects of annual distance driven and policy duration on claim count.

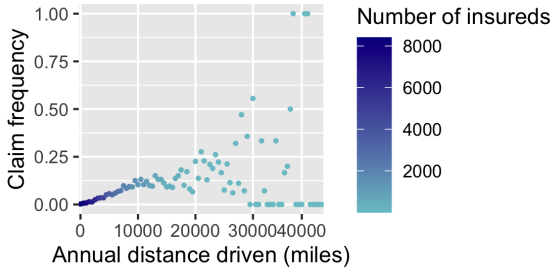

Figure 4.3: Distribution of claim frequency per annual distance driven

This section details the paper replication process where part 1 is used for independent cubic spline model for which annual distance driven and policy duration are fitted separately. Part 2 describes tensor product basis where the two variables are modelled together. The GLM model with policy duration as an offset is introduced in part 3 and the comparison of the three models is displayed in part 4. Finally, an extended GLM model with insured's age is added in part 5 and part 6 concludes the replication process with two modified GAM mimicking traditional GLM pricing structure.

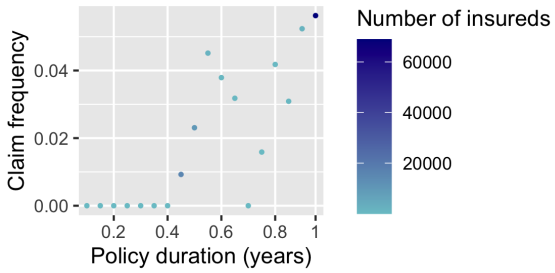

Figure 4.4: Distribution of claim frequency per policy duration

4.2.1 Independent Cubic Spline

In order to model annual distance driven and policy duration separately using cubic splines in a GAM setting, we have the model formula

$$log(y_i) = \beta_0 + f_1(distance_i) + f_2(duration_i) + \epsilon_i \tag{4.2.1}$$

where claim count is modelled with a Poisson distribution with log link function. $f_1(distance)$ and $f_2(duration)$ are cubic spline basis functions where non linear relationships between annual distance driven and claim frequency along with policy duration and claim frequency are captured independently. In order to better align with the original paper, basis dimensions of 7 and 3 are chosen respectively for annual distance driven and policy duration. Note that the knot locations are automatically chosen by `gam` function in package `mgcv` in R (Wood, 2017). Restricted Maximum Likelihood (REML) is used here in place of Generalized Cross Validation (GCV) in original paper for the smoothing parameter estimation due to its improved performance (Wood, 2011).

The modelling results of GAM using independent cubic splines can be seen from Table 4.3 where parametric term the intercept and the two non-parametric terms

Terms	Estimate	EDF	Significance
β_0	-3.715	-	< 0.001
$f_1(distance_i)$	-	4.530	< 0.001
$f_2(duration_i)$	-	1.975	< 0.001
REML score	11510		
AIC	22998.97		
BIC	23070.43		

Table 4.3: GAM using independent cubic splines results

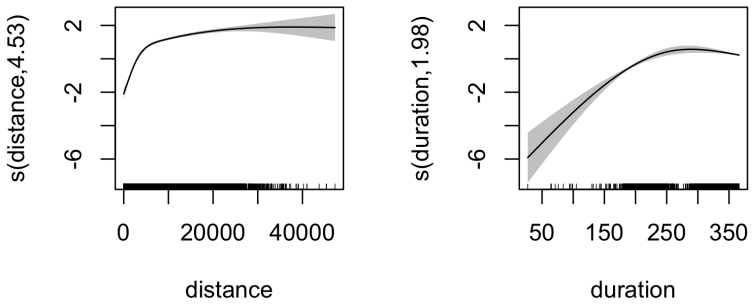

Figure 4.5: Partial effects of non-parametric terms in GAM with independent cubic spline

are all statistically significant with a p-value less than 0.001. REML score as well as AIC and BIC are also displayed for later model comparisons. Although the value of the intercept differs from the one in original paper (-2.735), they are of similar magnitude and both EDF (Effective Degrees of Freedom) of the two non-parametric functions matches the ones in original paper (4.30, 1.95).

For a more clear view at the effects of each predictor variable on the response claim count, we can take a look at partial effect plots (Figure 4.5) where the response is plotted against the fitted smooth terms.

One thing to note from Figure 4.5 is that the y-axis is in the scale of linear predictor but the results can be easily generalized to response level since exponential function is a strictly increasing function. Similar to *Figure 5* of original paper, as annual distance driven increases the effect of it on claim count also increases up till around 10000 miles where it reaches a maximum. As for the effect of policy duration, we observe a similar trend but with one difference where the effect of policy duration decreases after around 300 days. In addition, although the confidence interval is the widest around higher values of annual distance driven, the constant effect is significant and needs to be investigated further.

4.2.2 Tensor Product Base

In contrast to the independent cubic spline fit in the last section where two separate functions are used for annual distance driven and policy duration, modelling both variables together using tensor product base needs only a single function to account for individual variables effect as well as the added combined effect. That is

$$log(y_i) = \beta_0 + f(distance_i, duration_i) + \epsilon_i \tag{4.2.2}$$

where $f(distance_i, duration_i)$ allows the combined smoothing of two variables on different scales (Wood, 2017, Section 5.6). Again, the function is fitted with Poisson distribution using cubic splines with dimensions setting to 7 and 3 respectively for annual distance driven and policy duration.

Table 4.4 shows that both the intercept and the non-parametric term are significant at a significance level of < 0.001. The estimate of the intercept is very close to the estimate from the model fitted using independent cubic splines and this agrees with

Terms	Estimate	EDF	Significance
β_0	-3.722	-	< 0.001
$f(distance_i, duration_i)$	-	11.52	< 0.001
REML score	11505		
AIC	22988.88		
BIC	23109.29		

Table 4.4: GAM using tensor product base results

the original paper (*Table 5*). The GAM using tensor product base performs better than the GAM using independent cubic spline in terms of a lower REML score and a lower AIC. In addition, BIC of the tensor product model is higher than the one of the independent cubic spline model and this may be explained by the heavier penalty of BIC puts on more covariates.

In order to visualize the combined effect of annual distance driven and policy duration on claim frequency, we can look at 3-D perspective plots (Figure 4.6). Looking at the left figure which belongs to the independent cubic spline model, the yellow color indicates that insureds with a high annual distance driven and high policy duration combination have highest claims. However, the interpretation of the right figure is not that clear where claim frequency is fitted using tensor product. The part of the plot that is missing is the result of setting `too.far` to 0.08 which prompts the deletion of data points with predicted claims larger than 0.08 when predicted claim values are scaled to $0 \sim 1$. Without setting `too.far` argument in `vis.gam` the function plots all data points which will cause the z-axis of claim frequency reaches above 3 million which is absurd compared to actual maximum claim value of 3. This may indicate that the tensor product model has "considerable extrapolation" (S.Wood, personal communication, November 18, 2024) and as a result a lot of extreme data points are

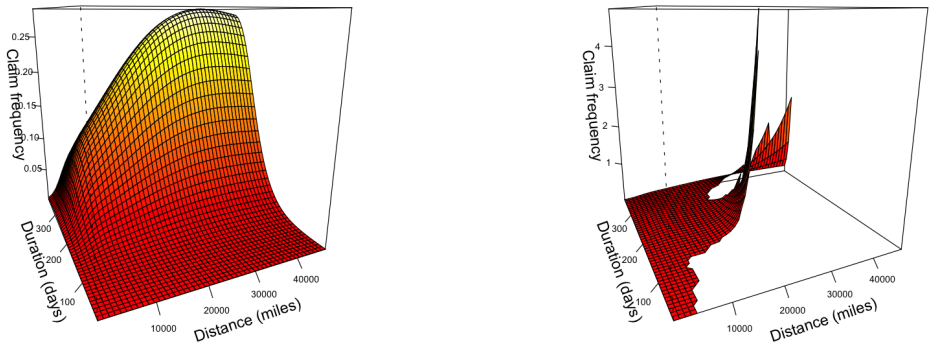

Figure 4.6: 3-D perspective plot visualizing the effects of annual distance driven and policy duration. The left figure belongs to independent cubic spline model and the right figure belongs to model with tensor product

generated upon prediction. Therefore, only the predicted data points that are close to range of true values are kept for visualization.

4.2.3 GLM with Offset

In order to approach the claim frequency prediction task like the traditional industry practice where GLM with Poisson distribution is fitted, we have the model construction

$$log(y_i) = \beta_0 + \beta_1 \cdot distance_{1i} + \beta_2 \cdot distance_{2i} + \beta_3 \cdot distance_{3i} + \beta_4 \cdot distance_{4i} + \beta_5 \cdot distance_{5i} + \log(duration_i) + \epsilon_i$$

where $\log(duration)$ is fitted as an offset and distance is converted into a factor variable with 6 levels where the base level corresponds to insureds with an annual distance driven between 1000 to 5000 miles. The distance variable has been cut with five splits to be comparable with the 7 knot locations of the previous GAM models. Table 4.5 gives detailed description of each annual distance driven covariate.

Covariate	Description
$distance_{1i}$	insureds with distance $0 \sim 1000$ miles
$distance_{2i}$	insureds with distance $5000 \sim 10000$ miles
$distance_{3i}$	insureds with distance $10000 \sim 15000$ miles
$distance_{4i}$	insureds with distance $15000 \sim 20000$ miles
$distance_{5i}$	insureds with distance $20000 \sim 47300$ miles

Table 4.5: Description of annual distance driven indicator variables

As it can be seen from Table 4.6, estimates of coefficients for annual distance driven are all above zero and have mostly an increasing trend except for the coefficient β_1 when distance is between 0 and 1000 miles. This is quite reasonable from intuition as normally we expect people who drive more to be exposed to more risks and therefore more likely to incur claims. All terms from the GLM model are significant at a significance level of 0.001 and the GLM model does produce a higher AIC and a higher BIC value comparing to both GAM models.

Figure 4.7 shows the 3-D perspective plot of the GLM model and we can see that it

Terms	Estimate	Significance
β_0	-9.466	< 0.001
β_1	-1.626	< 0.001
β_2	1.053	< 0.001
β_3	1.439	< 0.001
β_4	1.408	< 0.001
β_5	2	< 0.001
AIC	23308.07	
BIC	23363	

Table 4.6: GLM model results

is very similar to what we have for the GAM model with independent cubic splines (Figure 4.6) where the insureds with a high value of annual distance driven and a high value of policy duration have the largest risks.

4.2.4 Model Comparison Results

In this section, we compare the predictive results of the three models mentioned above using the data from validation set which contains 30% of the whole dataset. For all three models, MARs (Mean Absolute Residual) are computed and the distributions of absolute residuals across different actual claim count are shown in Figure 4.8.

We can see from Figure 4.8 that the three models exhibit similar distributions of absolute residuals and particularly when actual claim count is 0 both GAM models have distributions that is more leaned towards 0. When actual claim count is 1, GLM model displays a wider IQR (Inter Quartile Range) than other two models, which means GLM has more of a volatile distribution of absolute residuals which is similar to *Figure 14* of the original paper (Boucher et al., 2017). Although actual claim count spans values of 0, 1, 2, and 3, we decide not to show the distribution of

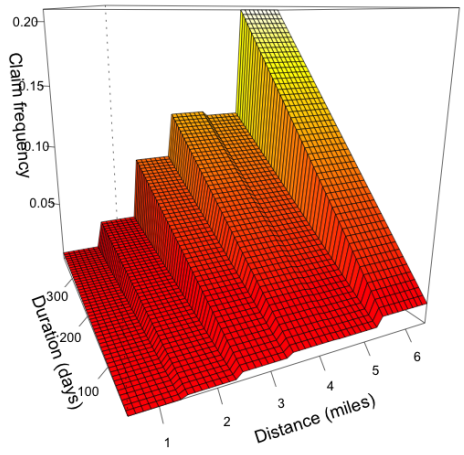

Figure 4.7: 3-D perspective plot of GLM model

residuals when claim count is 3 as there is only one data point under this situation in validation set (the proportions of data points with respect to actual claim count of 0, 1, 2 are: 95.6%, 4.2%, 0.2%). Furthermore, MAR results give us the worst model being the GLM model (MAR of 0.0843) and the two GAM models (MARs of 0.0841) display the same predictive power.

4.2.5 GLM with Insured's Age

Besides annual distance driven and policy duration, we can add other variables into our model such as insured's age, vehicle's age, insured's sex and type of parking which describe the drivers' characteristics and their driving behavior. Similar to *Section 3.2* in original paper where all of the above mentioned variables are modelled together under a GLM model for individual term significance analysis, we have a Poisson

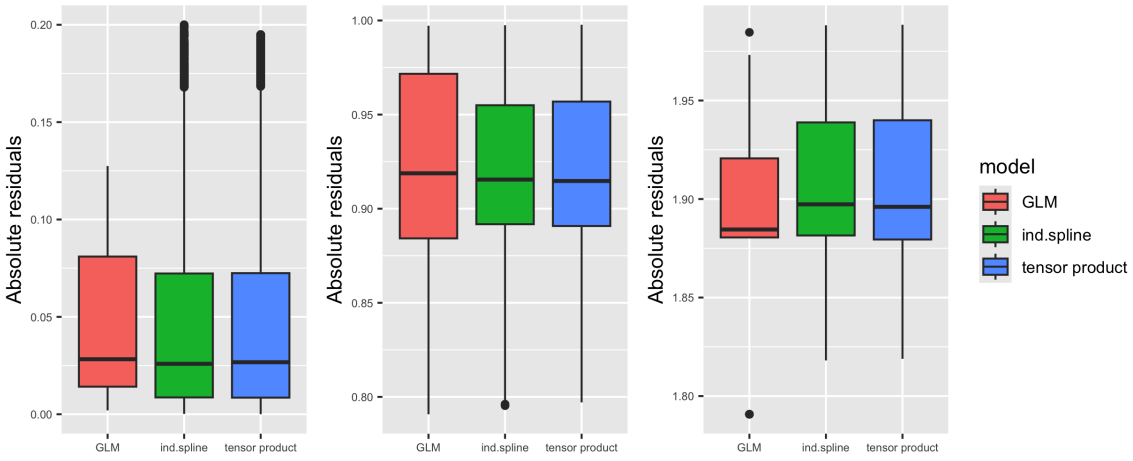

Figure 4.8: Distributions of absolute residuals across actual claim value 0, 1, 2 for all three models. From left to right, actual claim value of 0, 1, and 2

GLM with duration as an offset and other variables except the missing variable type of parking in the synthetic dataset with a log link structure. The modelling results show that both insured's age and vehicle's age are statistically significant (p-value < 0.01). Nonetheless, only insured's age are included for further modelling analysis where it is transformed into a factor variable with three levels (0-25, 25-30, 30-103) and it is added to all three models mentioned previously.

Therefore, we have the modified modelling formula for GAM with independent splines, GAM with tensor product, and GLM respectively

$$\log(y_i) = \beta_0 + f_1(distance_i) + f_2(duration_i) + \beta_1 \cdot age_{1i} + \beta_2 \cdot age_{2i} + \epsilon_i$$

and

$$\log(y_i) = \beta_0 + f(distance_i, duration_i) + \beta_1 \cdot age_{1i} + \beta_2 \cdot age_{2i} + \epsilon_i$$

and

$$\log(y_i) = \beta_0 + \beta_1 \cdot distance_{1i} + \beta_2 \cdot distance_{2i} + \beta_3 \cdot distance_{3i} + \beta_4 \cdot distance_{4i} + \beta_5 \cdot distance_{5i} + log(duration_i) + \beta_6 \cdot age_{1i} + \beta_7 \cdot age_{2i} + \epsilon_i$$

where age_{1i} and age_{2i} represent indicator variables of insured's age 0-25 and 25-30 (30-103 is set as the base group).

According to Table 4.7, we have very similar results with *Table 9* of the original paper where the estimates of insured's age are very close across three models and GAM with tensor product basis outperform GAM with independent splines and GLM in terms of lower REML and AIC values. In contradiction to the original paper, we do notice that the estimates of insured's age group 25-30 are insignificant for all three models which indicates the models suggest the risks of insured across 25-103 years old to be the same.

Terms	GAM(independent)	GAM(tensor)	GLM
age_{1i}	0.316	0.316	0.326
age_{2i}	-0.021*	-0.022*	0*
REML	11505	11500	11655
AIC	22986	22976	23294
BIC	23076	23115	23368

Table 4.7: Extended model result of the GAM with independent splines, GAM with tensor product, and GLM where insured's age is added to all three models. Values with * indicate insignificance at a level of 0.001.

As a result, we decide to change the cut points of insured's age to 25 and 40 such

that it has categories of 0-25, 25-40, and 40-103 by examining Figure A.1 where distribution of claim frequency with respect to insured's age are shown. Under this new transformation, the estimates of age'_{1i} and age'_{2i} become statistically significant with p-value < 0.05 across all three models (note: GLM age'_{2i} estimate has a p-value of 0.06).

Now we can report the estimated claim frequency for insureds across all possible age groups using a tariff structure. Table 4.8 gives the first part of the table where predicted claim frequency of all three models are recorded for insureds with age between 0 to 25 years. The second and third part of the table where predicted claim frequency for insureds with age 25-40, 40-103 can be found in Appendix B (Table B.1,Table B.2).

(distance,duration)	GAM(independent)	GAM(tensor)	GLM
(3500,0.35)	0.0042	0.0025	0.0135
(4500,0.5)	0.0279	0.0297	0.0192
(9000,0.65)	0.135	0.084	0.072
(15500,0.90)	0.231	0.151	0.142
(19000,1)	0.211	0.213	0.158

Table 4.8: Tariff structure of predicted claim frequency for insureds aged 0-25.

We can see from Table 4.8 that predicted claim frequency increases as annual distance driven and policy duration increase for all three models except for GAM with independent splines when insureds move from (15500,0.90) to (19000,1). This decrease in predicted claim frequency could be explained by looking at 3-D perspective plot (Figure A.3) where we can observe a gradual decrease in claim frequency from a distance-duration combination of (15500,0.90) to (19000,1) as we move from the ridge of the curved surface to upper left direction. Another more intuitive explanation of

this might be that the insured who drive above certain annual distance as well as obtain a high policy coverage would experience practice effect. As a result, they get more good at driving and hence are less risky to obtain a claim.

4.2.6 GAM Pricing Structure

In this last section of the paper replication process, we introduce an insurance pricing structure with GAM which mimics a traditional GLM pricing structure. Similar to *Section 3.3* of the original paper, two modified GAMs are introduced where knot numbers are increased with knots locations fine-tuned such that we have a detailed risk segmentation for each insured. To construct the GAM models, we still have the same Poisson GAMs with independent cubic splines and tensor product but knot locations at each 500 miles for annual distance driven and 0.05 years for policy duration. In order to estimate the potential claim frequency of an insured, we first calculate the linear predictor effects from both annual distance driven and policy duration. Next, we exponentiate the added total effect (total effect = distance effect + duration effect) plus intercept and arrive at the response level effect. Looking at Table 4.9, the fourth column Total Relativity equals to the product of the second and the third column which are the separated effects of distance and duration respectively.

Table 4.9 illustrates the tariff insurance ratemaking structure under a GAM model using independent cubic splines. We can observe that from second column as annual distance driven increases, the effect of it on claim frequency also increases monotonically which agrees with the results from *Table 11* of original paper. Same for policy duration in column 3 where monotonically increasing effects on claim frequency are shown.

(distance,duration)	Rela.(distance)	Rela.(duration)	Total Rela.	Claim Freq.
(3500,0.35)	1.27	1.56×10^{-5}	1.99×10^{-5}	4.64×10^{-7}
(4500,0.5)	1.70	0.4	0.68	0.016
(9000,0.65)	3.40	1	3.38	0.079
(15500,0.90)	3.60	1.25	4.49	0.10
(19000,1)	4.61	1.34	6.16	0.14

Table 4.9: Tariff structure for modified GAM with independent splines where annual distance driven has 95 knots at every 500 miles and policy duration has 21 knots at every 0.05 years.

Similarly, we can modify GAM with tensor product to produce such tariff structure. Table 4.10 gives us the tariff structure under a GAM with tensor product where knot locations are at every 500 miles for annual distance driven and every 0.05 years for policy duration. Unlike *Table 12* of original paper where monotonically increasing patterns are observed for the combined effects of distance and duration, Table 4.10 displays a mostly increasing pattern until a change from (distance,duration) combination of (15500,0.90) to (19000,1). This pattern could be verified by looking at the predicted surface of the GAM with tensor product base at a different angle. (Figure A.6)

(distance,duration)	Rela.(distance,duration)	Claim Freq.
(3500,0.35)	4.04×10^{-5}	9.29×10^{-7}
(4500,0.5)	0.80	0.018
(9000,0.65)	2.25	0.052
(15500,0.90)	10.5	0.24
(19000,1)	6.16	0.14

Table 4.10: Tariff structure for modified GAM with tensor product where annual distance driven have knots at every 500 miles and policy duration has knots at every 0.05 years.

Figure 4.9 gives us the predicted surface for both GAM models. Both figures resemble

Figure 4.9: 3-D perspective plots of the modified GAM with knot locations at every 500 miles for annual distance driven and 0.05 years for policy duration. The left figure is from GAM with independent cubic splines and the right figure is from GAM with tensor product.

the shapes from Figure 4.6 where predicted surface for original GAMs are depicted. One improvement is that we are seeing a more reasonable scale on z-axis of claim frequency where a maximum value of just over 20 claims are observed whereas a value of 3 million are seen in Figure 4.6 and this may due to the effects being smoothed out for each (distance,duration) category and thus are less prone to extreme extrapolations given the added flexibility of the tensor product base.

4.2.7 Comparisons and Remarks

Through our analysis, we have found that although normally a linear effect of policy duration on claim frequency is assumed in insurance practice, it is often inappropriate as our model shows that a diminishing effect is seen as policy duration saturates that is we do not observe a doubling in risks as policy duration of an insured doubles.

Similarly, insureds who drive twice the distance do not indicate a doubling in risks. Annual distance driven seems to have a decreasing effect on claim frequency that is as more distance travelled by an insured the risks increase more slowly which agrees with Boucher et al. (2017).

Most of our data analysis results agree with the results in Boucher et al. (2017) which is as expected since we are performing an analysis using a synthetic dataset that is generated on purpose to mimic the original dataset with the aim to protect privacy of insureds. However, some of our results differ from the results in original paper and that is probably due to the difference in data structure of the synthetic dataset from the original data set. And this could be seen from the different distribution of policy duration (Figure 4.2) that is potentially caused by data generating algorithm used. In addition, the decreasing effects patterns that could be seen in Table 4.8 and Table 4.10 as one moves from a lower (distance,duration) combination to a higher segment could be explained by looking at 3-D perspective plots where predicted frequency surfaces are displayed (Figure A.3, Figure A.6). This effect could be explained by the combined influence of both annual distance driven and policy duration. This trend can only be observed with drivers who drive above a certain (distance,duration) combinations which means that drivers who drive above a certain number of annual distance and have a policy duration longer than a certain time would experience practice effects where they bear a lower risks than drivers who drive less and have a lower policy duration.

According to section 4.2.4 where two GAMs are compared with GLM, the results from two GAMs are very similar to the results from GLM. In particular, the GLM obtains a close value of median absolute residuals as the two GAMs while producing a more

volatile residuals distribution when actual claims are 1. Nonetheless, we still think GLM model is the better model in terms of its ease of interpretation and simplicity of model construction where no knots need to be chosen.

However, the trade-off between interpretability and predictability becomes unbalanced if one can find a model which produces more accurate predictions than the GLM model. Therefore, we attempt to tackle the challenge with two approaches in the next two sections. One way is to add more telematics variables which hopefully will give us more insights regarding the driving behavior of the insured and thus might improve claim frequency prediction. The other method is to introduce a robust alternative to the GAM models where individual data point are evaluated and assessed for their influence on the whole model thus reducing effects from outliers which may also improve claim prediction accuracy.

4.3 Model Extensions

4.3.1 Addition of More Telematics Variables

In order to come up with which variable(s) to add to our existing models, we perform machine learning methods to evaluate individual variable importance in all 45 continuous variables. Particularly, we employ two different machine learning methods: Random Forest and Extreme Gradient Boosting (Xgboost) with Multivariate Adaptive Regression Splines (MARS) using `train()` in package `caret` in R (Kuhn and Max, 2008). Random Forest and Xgboost are chosen for their popularity within the industry and MARS is chosen for its close relationship with GAM.

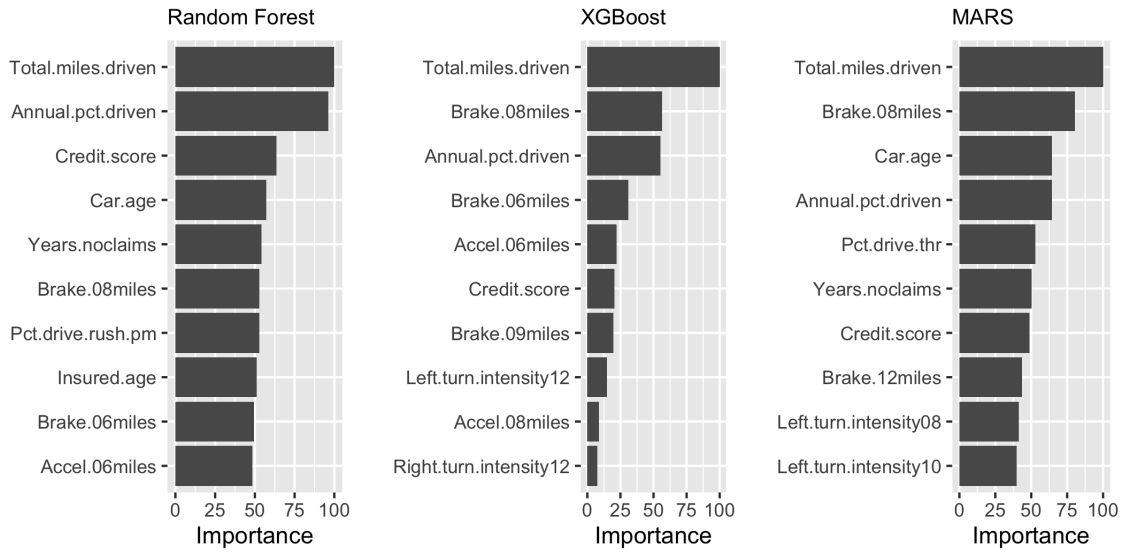

Figure 4.10: Variable importance plots produced by three models. From left to right, Random Forest, XGBoost, MARS. Variables importance are evaluated on RSS of claim count reduced.

All three models are tuned so that Root Mean Squared Error (RMSE) are minimized using 5 fold cross-validation. The idea is to find the variables that predict claim count most accurately in terms of largest reduction in Residual Sum of Squares (RSS). Looking at Figure 4.10, we can observe that variable `Brake.08miles` is included in all three plots and it ranks the second just after annual distance driven for both XGBoost and MARS. Upon close examination, `Brake.08miles` represents the telematics information of number of sudden brakes 8 mph/s per 1000 miles and it displays an increasing relationship with claim frequency in the partial dependence plots produced by all three machine learning methods (Figure A.4).

Therefore, we decide to add `Brake.08miles` as an additional variable in our original GAM models. Looking at Figure A.2 where distribution of claim frequency with respect to number of sudden brakes 8 mph/s per 1000 miles is shown, one can clearly see that number of sudden brakes has a positive relationship with claim count around

Terms	Estimate	EDF	Significance
β_0	-3.827	-	< 0.001
$f_1(distance_i)$	-	4.604	< 0.001
$f_2(duration_i)$	-	1.963	< 0.001
$f_3(brakes_i)$	-	1.978	< 0.001
REML score	11245		
AIC	22469.19		
BIC	22560.22		

Table 4.11: Three-variable GAM using independent cubic splines results

values where data density is high.

Fitting the Poisson GAM with independent splines while adding the third variable brakes is mostly the same as we did above except we are estimating three smoothed functions now instead of two previously. The model fitted with REML is structured as below:

$$log(y_i) = \beta_0 + f_1(distance_i) + f_2(duration_i) + f_3(brakes_i) + \epsilon_i$$

where $f_1(distance_i)$ and $f_2(duration_i)$ are fitted using cubic splines with number of knots 7 and 3 just like before. $f_3(brakes_i)$ is fitted using cubic splines with 3 knots since 3 is the minimal knots needed for the brakes variable (`k.check()` shows sufficient basis dimensions).

Table 4.11 gives us a summary of the results obtained after fitting the three-variable GAM model using independent cubic splines. As we can observe, all three smoothed functions are statistically significant with p-values less than 0.001. Comparing to Table 4.3, our three-variable GAM with independent cubic splines produces similar estimates but performs better than two-variable GAM across all three metrics REML

Terms	Estimate	EDF	Significance
β_0	-3.787	-	< 0.001
$f(distance_i, duration_i, brakes_i)$	-	23.79	< 0.001
REML score	11195		
AIC	22383.77		
BIC	22629.85		

Table 4.12: Three-variable GAM using tensor product base results

score, AIC, and BIC. The partial effects plot (Figure A.5) of variable brakes on claim frequency further confirms previous observation that an increasing relationship is present between the two variables.

Similarly, we can add variable brakes into the GAM with tensor product base along with annual distance driven and policy duration. The model structure would then be

$$log(y_i) = \beta_0 + f(distance_i, duration_i, brakes_i) + \epsilon_i$$

where $f(distance_i, duration_i, brakes_i)$ is fitted using tensor product base and basis dimensions are set to 7, 3, 3 for annual distance driven, policy duration, and number of sudden brakes respectively.

Table 4.12 gives us the results after fitting three-variable GAM with tensor product base. Again, just like the original GAM with tensor product base, this model is fitted using Poisson distribution with log link and REML. Looking at Table 4.12, we can observe that the three-variable GAM with tensor product base produces very similar estimates with original two-variable GAM (Table 4.4). Moreover, our three-variable GAM outperforms two-variable GAM in terms of a lower REML score, AIC, and BIC. Even though we are seeing an improvement in performance across REML score, AIC, and BIC which are all model based evaluation methods, we are more interested in the

predictive performance of our models in practice. That is, how good are our models in predicting the unseen data through non model based evaluation metrics. One way to tell is by looking at MAR across all models where MAR has the formula

$$MAR = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

where y_i are the actual claim count and $\hat{y}_i$ are the predicted claim count. MAR is a metric that is very similar to RMSE (Root Mean Squared Error) for the reasons that they are both of data based methods and they are giving us an idea of how close our predictions are compared to true values on average.

Two-variable Models	GAM(ind.)	GAM(tensor)	GLM*
MAR	0.08409	0.08410	0.08428
Three-variable Models	GAM(ind.)	GAM(tensor)	GLM*
MAR	0.08249	0.08208	0.08392

Table 4.13: MAR results across GAM with independent splines, tensor product base and GLM with factor variables for two-variable and three-variable models. * indicates that the model is fitted with factor variables using cut rules correspond to respective basis dimensions in GAM.

We can now observe from Table 4.13 that our three-variable models outperform all two-variable models with same model structures. Particularly, three-variable GAM with tensor product base is the best model overall in terms of having the lowest MAR. As one may notice, the three-variable GLM model is also reported in the table and we can see that it has improved over the two-variable GLM model.

In conclusion, with the help of machine learning methods, the addition of another telematics variable brakes makes our models better across the board. With the added

brakes variable, our GAM models improve significantly from their two-variable predecessors and the GLM model improves as well. However, just like we have touched above where Figure 4.8 shows similarity in distributions between GLM and GAMs but existence of extreme extrapolations in GAMs brings concerns. In the next section, we will try to solve this challenge by incorporating robust GAM models where influences of outliers may be weakened.

4.3.2 A Robust GAM Approach

In this section, we will apply robust extension of GAM using function `gamlss()` developed by Aeberhard et al. (2021) in package `GJRM` in R (Marra and Radice, 2023).

Similar to previous sections, firstly we fit the two-variable robust GAM with Poisson distribution where claim count is modeled with distance driven and policy duration. The modeling structure is the same as 4.2.1 for an independent cubic spline model. As we can observe from Table 4.14, the two-variable robust model shares similar intercept estimate with non-robust counterpart in Table 4.3 but has a lower effective degrees of freedom. This tells us that as expected the robust GAM automatically detect outlying points and reduce their influence to model fit as a result the smoothed terms are less flexible comparing to non-robust GAM.

Next, the three-variable robust GAM is fitted and Table 4.14 shows a similar pattern with above two-variable robust model where it exhibits a similar coefficient estimate but shows a lower effective degrees of freedom comparing to Table 4.11. Two-variable and three-variable tensor product base is also fitted with robust GAM where they share similar results with their non-robust counterparts in Table B.3 in Appendix.

Two variables model			
Terms	Estimate	EDF	Significance
β_0	-3.680	-	< 0.001
$f_1(distance_i)$	-	2.270	< 0.001
$f_2(duration_i)$	-	0.503	< 0.001
Three variables model			
β_0	-3.941	-	< 0.001
$f_1(distance_i)$	-	2.478	< 0.001
$f_2(duration_i)$	-	0.495	< 0.001
$f_3(brakes_i)$	-	1.554	< 0.001

Table 4.14: Robust GAM using independent cubic splines results

Looking at Table 4.15, we observe a similar pattern as the one in Table 4.13 where in the two-variable scenario robust GAM with tensor product base performs slightly worse than robust GAM with independent spline basis. However, the three-variable robust GAM improves significantly from the two-variable models and all of them show improvements over their non-robust counterparts by having lower MAR's. This result validates our hypothesis earlier that the robustness property of the robust GAM can be taken advantage of for better model prediction. The 3-D perspective plots of both models can be found in Appendix A.7.

Two-variable Models	robust GAM(ind.)	robust GAM(tensor)
MAR	0.08370	0.08380
Three-variable Models	robust GAM(ind.)	robust GAM(tensor)
MAR	0.08203	0.08186

Table 4.15: MAR results for robust GAM with both two variables and three variables scenarios. All are fit with Poisson distribution with log link and extended Fellner-Schall method for smoothing parameter selection.

	MAR			MAR		
Outliers(%)	GAM Indp.	rGAM Indp.	Δ%	GAM Ten.	rGAM Ten.	Δ%
0%	0.0841	0.0837	−0.5	0.0841	0.0838	−0.36
1%	0.1784	0.1334	−25	0.1783	0.1335	−25
2%	0.2701	0.1829	−32	0.2700	0.1828	−32
3%	0.3604	0.2328	−35	0.3603	0.2329	−35

Table 4.16: Comparison of MAR produced by two-variable GAM and robust GAM as outliers increase from 0% to 3%.

However, since the maximum claim count is 3 in our data, we decide to manually generate outliers to further compare model performances between robust and non-robust GAM's. The outliers are firstly drawn randomly via stratified cross validation where a preset percentage of outliers are chosen randomly from each training data and validation data. This ensures that the proportions of outliers from both the training and validation data match the predefined threshold. For generality we simulate outliers of 0% to 3% with a step size of 1% where 0% represent the original unadjusted data. The chosen data points are then enlarged in response level by adding 5 if the true value is 0 and enlarged by 6 times if the true value is larger than 0. Therefore, we would have outliers of claim count values of 5, 6, 12, 18 comparing to original values of 0, 1, 2, 3. For verification, package `DHARMa` is used for bootstrapping and testing if the data has more outliers than expected under the assumed Poisson model (Hartig, 2022). The results are perfectly in agreement with our intention for which it shows that the test is insignificant under original data meaning the original data has normal amount of outliers and does not deviate from model assumptions. However, as we increase the amount of outliers from 1% to 3% the tests are significant and correctly output the matching outliers proportions of 1% to 3% in Table B.4.

Looking at Table 4.16, we observe that although robust GAM with tensor product

	Lift		Lift	
Outliers(%)	GAM Indp.	rGAM Indp.	GAM Ten.	rGAM Ten.
0%	42.03	31.54	44.10	29.23
1%	3.85	30.29	3.75	29.22
2%	2.43	28.71	2.49	28.82
3%	1.85	26.91	1.95	20.90

Table 4.17: Comparison of Lift produced by two-variable GAM and robust GAM as outliers increase from 0% to 3%.

basis performs similarly to robust GAM with independent cubic splines, they both improve significantly from non-robust GAM as outliers introduced increase from 0% to 3%. The increasing $\Delta\%$ improvements indicate that as we increase amount of outliers in the data the non-robust models become distorted leading to a poor fit whereas the robust models are able to control the effects of outliers with a bounded influence function thus would produce a more robust fit.

The robust models also improve upon data segmentation. From Table 4.17, it can be concluded that when there is no outliers the robust models perform similarly with non-robust counterparts by having a similar Lift where Lift is a measure devised for measurement of extremeness between data points and it can be calculated as

$$\text{Lift} = \frac{Mean(\text{Top 10\% predicted})}{Mean(\text{Lower 10\% predicted})} \tag{4.3.1}$$

(Meng et al., 2022). Therefore, the higher the Lift the better the ability of the model discriminating extreme points. As we increase added outliers, the robust model is able to retain a similar level of separation among predicted values but non-robust counterparts quickly deteriorate producing a much lower Lift. When combined with results from MAR, we found that the robust GAM's are able to relax model assumptions

due to its robustness nature and as a result more accurate predictions and data segmentation are produced. A similar pattern can also be observed from three-variable comparisons (see Table B.5 and B.6).

To conclude, in situations where small to intermediate outliers are present which is usually the case in practice, actuaries can utilize the advantage of robustness of robust GAM to produce a more accurate model prediction. The improvement in prediction of robust GAM to non-robust models can also be seen from an improvement of data segmentation where the model is better able to identify extreme points and thus risks are better identified for each insured.

Chapter 5

Conclusion

As modern technological developments flourish, actuaries in P&C now are able to get hands on variety of driving behavior data for each policyholder through telematics. This huge amount of data possesses potential for better risk classification but at the same time introduces challenges of effective usage. To leverage this, GAM and GAM with machine learning variable filtering can be employed for a more flexible fitting and more accurate predictions can be produced comparing to standard GLM. GAM offers flexible modeling while maintaining a level of interpretation which is seen as essential to actuarial rate regulation.

Robust GAM improves further on model prediction and risk segmentation with the help of its influence measurement of outlying points. Automatic detection of outliers gives robust GAM the ability of down-weighing influential points. As a result of this, it grants us a more accurate model and a robust alternative in the presence of extreme outliers.

For future studies, more research can be focused on identifying potential outlying insureds and combining claim count prediction with claim severity prediction where

they are often observed to be correlated. This may be done by developing a GAM with dependent structure between claim frequency and claim severity through either incorporating dependence in a pure premium rating structure using for example copulas or by adding claim count as a covariate for modeling claim severity in a robust GAM setting where the effects of often observed outliers in skewed distribution of claim severity could be alleviated.

Another task is to simplify modeling process for both GAM and robust GAM in insurance pricing. Although GAM and robust GAM beat traditional GLM with higher model accuracy, the fitting process is quite involved due to basis dimension and knots location selection which often needs expert opinions. Other studies could be focused on model selection in P&C ratemaking in general where it is seen as a fast growing industry with emergence of A.I. in recent years. There have been discussions and debates in the actuarial community regarding practicality of employing sophisticated pricing models in practice because of long development time with such models. On the one hand, we are seeing more and more companies moving towards an interactive platform where software does model tuning and the heavy lifting part and hence saving time and efforts for the actuaries. Whereas majority companies still face this dilemma where the trade-off between sticking with the traditional pricing modeling procedure such as GLM and investing more on interactive platforms seems unclear. With developments such as driver assistance features and automated driving getting popular, insurance pricing industry will face another challenge that could potentially shift the focus of the ratemaking process.

Appendix A

Figures

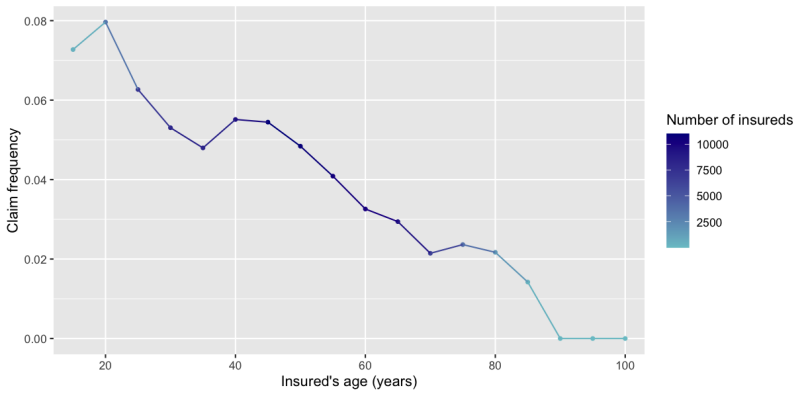

Figure A.1: Distribution of claim frequency with respect to insured's age.

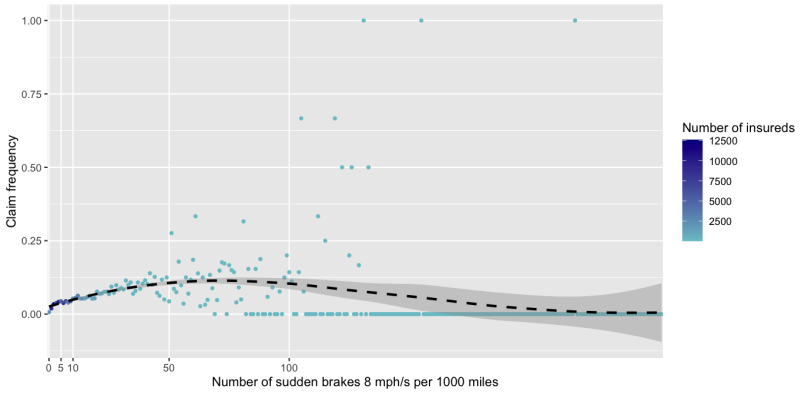

Figure A.2: Distribution of claim frequency with respect to number of sudden brakes 8mph/s per 1000 miles.

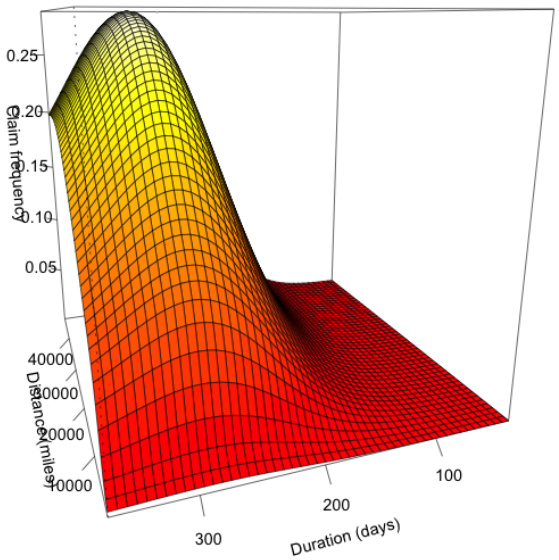

Figure A.3: 3-D perspective plot of GAM independent cubic spline with insured's age added as a factor variable visualizing the effects of annual distance driven and policy duration.

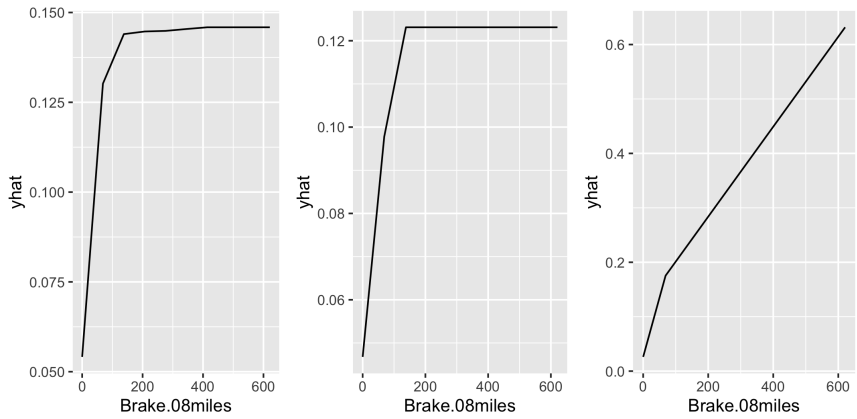

Figure A.4: Partial dependence plots of effects of Brake.08miles on claim frequency. The vertical axis represents predicted claim count. From left to right: Random Forest, XGBoost, MARS.

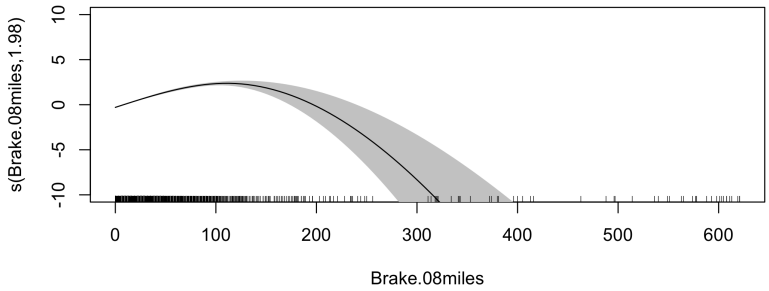

Figure A.5: Partial effects plot of Brakes.08miles of three-variable GAM model. The scale is on linear predictors and the limits on y-axis has been trimmed down to (-10,10).

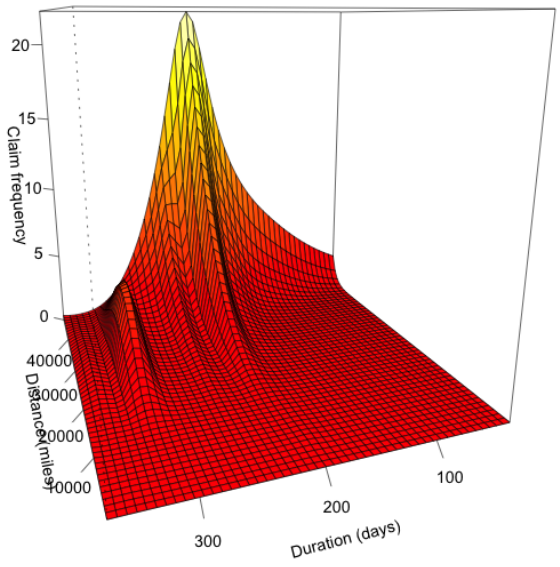

Figure A.6: 3-D perspective plot of modified GAM with tensor product base displaying the effects of annual distance driven and policy duration on claim frequency. Knots are placed at every 500 miles for annual distance driven and every 0.05 years for policy duration.

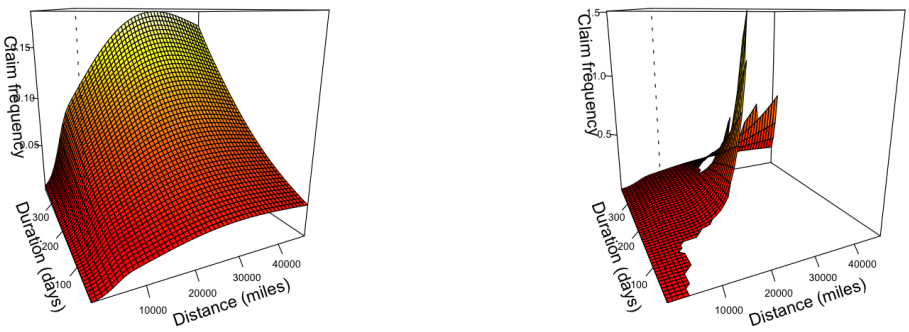

Figure A.7: 3-D perspective plot of robust GAM fitting claim count on distance driven and policy duration. The left figure is independent cubic spline model and the right figure is robust tensor product.

Appendix B

Tables

(distance,duration)	GAM(independent)	GAM(tensor)	GLM
(3500,0.35)	0.0028	0.0017	0.0092
(4500,0.5)	0.019	0.0200	0.0132
(9000,0.65)	0.0915	0.0564	0.0493
(15500,0.90)	0.156	0.102	0.097
(19000,1)	0.142	0.144	0.108

Table B.1: Tariff structure of predicted claim frequency for insureds aged 25-40.

(distance,duration)	GAM(independent)	GAM(tensor)	GLM
(3500,0.35)	0.003	0.0019	0.01
(4500,0.5)	0.021	0.022	0.014
(9000,0.65)	0.102	0.063	0.053
(15500,0.90)	0.174	0.114	0.105
(19000,1)	0.159	0.161	0.117

Table B.2: Tariff structure of predicted claim frequency for insureds aged 40-103.

Terms	Estimate	EDF	Significance
Two variable model			
β_0	-3.665	-	< 0.001
$f(distance_i, duration_i)$	-	5.172	< 0.001
Three variable model			
β_0	-4.795	-	< 0.001
$f(distance_i, duration_i, brakes_i)$	-	25.05	< 0.05

Table B.3: Robust GAM using tensor product base results

Outliers	GAM Poi Ind.	GAM Poi Tens.
0%	0.0013, 0.99	0.0013, 1
1%	0.01, <0.01	0.01, <0.01
2%	0.02, <0.01	0.02, <0.01
3%	0.03, <0.01	0.03, <0.01

Table B.4: Outlier frequency and p-value for outliers test. Models are Poisson GAM with independent splines and Poisson GAM with tensor product base. The p-values are based on test of outliers produced by `testOutliers` in R package `DHARMa`. The hypothesis test is to use bootstrap to simulate fitted values from the specified model and to test whether observed data has more values that is larger than expected setting `alternative="greater"`, `margin="upper"`, `type="bootstrap"`.

	MAR			MAR		
Outliers(%)	GAM Indp.	rGAM Indp.	Δ%	GAM Ten.	rGAM Ten.	Δ%
0%	0.0825	0.0820	-0.6	0.0821	0.0818	-0.4
1%	0.1775	0.1318	-26	0.1765	0.1320	-25
2%	0.2693	0.1812	-33	0.2684	0.1811	-33
3%	0.3597	0.2311	-36	0.3588	0.2319	-35

Table B.5: Comparison of MAR produced by three-variable GAM and robust GAM as outliers increase from 0% to 3%.

	Lift		Lift	
Outliers(%)	GAM Indp.	rGAM Indp.	GAM Ten.	rGAM Ten.
0%	63.87	71.02	56.91	121.06
1%	4.54	67.42	4.39	43.20
2%	2.74	63.04	2.75	103.45
3%	2.06	59.23	2.13	44.52

Table B.6: Comparison of Lift produced by three-variable GAM and robust GAM as outliers increase from 0% to 3%.

Bibliography

William H. Aeberhard, Eva Cantoni, Giampiero Marra, and Rosalba Radice. Robust fitting for generalized additive models for location, scale and shape. 31(1):11, 2021. doi: 10.1007/s11222-020-09979-x.

Jean-Philippe Boucher, Steven Côté, and Montserrat Guillen. Exposure as duration and distance in telematics motor insurance using generalized additive models. 5 (4):54, 2017. doi: 10.3390/risks5040054. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

George Casella and Roger L. Berger. Statistical inference, 2001. URL https://www.routledge.com/Statistical-Inference/Casella-Berger/p/book/9781032593036.

Le Chang, Guangyuan Gao, and Yanlin Shi. Claims reserving with a robust generalized additive model. 28(4):840–860, 2024. doi: 10.1080/10920277.2023.2259445.

Tingting Chen, Anthony Francis Desmond, and Peter Adamic. Generalized additive modelling of dependent frequency and severity distributions for aggregate claims. pages 1–37, 2023. doi: 10.47260/jsem/1241.

Claudia Czado, Julia Pfettner, Susanne Gschloßl, and Frank Schiller. Nonnested

model comparison of GLM and GAM count regression models for life insurance data. 2009.

Michel Denuit and Stefan Lang. Non-life rate-making with Bayesian GAMs. 35(3): 627–647, 2004. doi: 10.1016/j.insmatheco.2004.08.001.

Zuleyka Díaz Martínez, José Fernández Menéndez, and Luis Javier García Villalba. Tariff analysis in automobile insurance: Is it time to switch from generalized linear models to generalized additive models? 11(18):3906, 2023. doi: 10.3390/math11183906. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

Ludwig Fahrmeir, Stefan Lang, and Friedemann Spies. Generalized geoadditive models for insurance claims data. 26(1):7–23, 2003. doi: 10.1007/BF02808770.

Mark Goldburd, Anand Khare, Dan Tevet, and Dmitriy Guller. *Generalized Linear Models for Insurance Rating (Second Edition)*. Casualty Actuarial Society, 2016.

Florian Hartig. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*, 2022. URL `https://CRAN.R-project.org/package=DHARMa`. R package version 0.4.6.

Trevor Hastie and Robert Tibshirani. *Generalized Additive Models*. Chapman and Hall/CRC, 1986.

Kivan Kaivanipour. *Non-Life Insurance Pricing Using the Generalized Additive Model, Smoothing Splines and L-Curves*. 2015. URL `https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-168389`.

Nadja Klein, Michel Denuit, Stefan Lang, and Thomas Kneib. Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. 55:225–249, 2014. doi: 10.1016/j.insmatheco.2014.02.001.

Kuhn and Max. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. doi: 10.18637/jss.v028.i05. URL `https://www.jstatsoft.org/index.php/jss/article/view/v028i05`.

Giampiero Marra and Rosalba Radice. *GJRM: Generalised Joint Regression Modelling*, 2023.

Lih-Wen Mau, Jaime M. Preussler, Linda J. Burns, Susan Leppke, Navneet S. Majhail, Christa L. Meyer, Tatenda Mupfudze, Wael Saber, Patricia Steinert, and David J. Vanness. Healthcare costs of treating privately insured patients with acute myeloid leukemia in the United States from 2004 to 2014: A generalized additive modeling approach. 38(5):515–526, 2020. doi: 10.1007/s40273-020-00891-w.

Shengwang Meng, Yaqian Gao, and Yifan Huang. Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees. 106:115–127, 2022. doi: 10.1016/j.insmatheco.2022.06.001.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. 135(3):370, 1972. doi: 10.2307/2344614.

Fevi Novkaniza, Alfina Wijaya, and Sindy Devila. Generalized additive model for vehicle insurance premium calculation based on mileage and contract duration. 2025.

David Pitt, Stefan Trück, Rob van den Honert, and Wan Wah Wong. Modeling risks from natural hazards with generalized additive models for location, scale and shape. 275:111075, 2020. doi: 10.1016/j.jenvman.2020.111075.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. 54(3):507–554, 2005. doi: 10.1111/j.1467-9876.2005.00510.x.

Mitchell Roznik, C. Brock Porth, Lysa Porth, Milton Boyd, and Katerina Roznik. Improving agricultural microinsurance by applying universal kriging and generalised additive models for interpolation of mean daily temperature. 44(3):446–480, 2019. doi: 10.1057/s41288-019-00127-9.

Eslam Seyam and Elsalmouny Hussien. Proposed models for comprehensive automobile insurance ratemaking in Egypt with parametric and semi-parametric regression: A case study. 11(1):41–55, 2022. doi: 10.18576/jsap/110104.

Banghee So, Jean-Philippe Boucher, and Emiliano A. Valdez. Synthetic dataset generation of driver telematics. 9(4):58, 2021. doi: 10.3390/risks9040058. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Chen Tingting. Generalized additive models for dependent frequency and severity of insurance claims, 2018. URL `http://hdl.handle.net/10214/14769`.

Simon N. Wood. Fast stable restricted maximum likelihood and marginal likelihood

estimation of semiparametric generalized linear models. 73(1):3–36, 2011. doi: 10.1111/j.1467-9868.2010.00749.x.

Simon N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman and Hall/CRC, 2 edition, 2017. doi: 10.1201/9781315370279.

Shengkun Xie. Estimating risk relativity of driving records using generalized additive models: A statistical approach for auto insurance rate regulation. 18(1):55–86, 2024. doi: 10.1515/apjri-2023-0032. Publisher: De Gruyter.

Shengkun Xie and Kun Shi. Generalised additive modelling of auto insurance data with territory design: A rate regulation perspective. 11(2):334, 2023. doi: 10.3390/math11020334. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.