

# GLM, GNM and GAM Approaches on MTPL Pricing

Claudio Giorgio Giancaterino

*Aviva Italia Servizi S.c.a.r.l..*

## Abstract

This work has the aim to compare different statistics modelling approaches both used and could be second-hand used in motor third party liability pricing.

Generalized Linear Models are made up by a systematic component  $\eta_i = \sum_{j=1}^n x_{ij}\beta_j$  linked with a random component  $Y_i = EF(b(\theta_i); \frac{\phi_i}{\omega_i})$  by a link function  $g(\mu_i)$ . A Generalized Non-linear Model is the same as the GLM model except the link function  $g(\mu_i) = \eta_i(x_{ij}; \beta_j)$  where the systematic component is non linear in the parameters  $\beta_j$ . Generalized Additive Models extend Generalized Linear Models in the predictor  $\eta_i = \sum_{p=1}^n x_{ip}\beta_p + \sum_{j=1}^n f_j(x_{ij})$  made up by one parametric linear part and one non parametric part built by the sum of unknown "smoothing" functions of the covariates.

		Mean commercial tariff	Tariff requirement	Loss Ratio	Residuals degrees of freedom	Expected Losses	Actual Losses	Explained Deviance	Risk coefficients
Uni-Variate Analysis	<b>GLM</b>	234,4587	1,000490	1,447822	27.501	9.337.547	9.342.125	96,96%	20
	<b>GNM</b>	234,4647	1,000476	1,447785	27.501	9.337.683	9.342.125	96,96%	20
	<b>GAM</b>	232,8702	1,001729	1,457698	27.476	9.325.999	9.342.125	96,20%	45
Multi-Variate Analysis	<b>GLM</b>	234,6486	0,9981246	1,446650	27.505	9.359.678	9.342.125	87,64%	16
	<b>GNM</b>	234,6165	0,9979703	1,446848	27.505	9.361.125	9.342.125	87,04%	16
	<b>GAM</b>	248,5732	0,8596438	1,365612	27.265	10.867.438	9.342.125	84,80%	256

GAM approach is flexible to fit data, with realistic values and low level of residual deviance, but quite complex to realize. GLM is easier to use, but sometimes with overestimated coefficients and high values about residual deviance. GNM is an upgrade of GLM model, it grants some elaborations that

GLM can't replicate and with lower values compared to it.

From these three models GAM is able to personalize a premium with more risk coefficients.

*Keywords:* Actuarial Sciences, Non-Life Insurance Pricing, Statistical Modelling.

## Introduction

This work has the aim to compare different statistics modelling both used and could be used in motor third party liabilities.

The reason why is the continuous growing up of statistical methodologies, but GLM remain the most popular model used for GI development products. Rising up needs to customize the non-life insurance premium carry on looking for a model can replace GLM. In this paper are compared both classical GLM with GAM and with GNM approach, recently discovered (2007-2008).

## Assumptions

The dataset used in this working paper is the “ausprivauto0405” data frame bundled within R “CASDatasets” package. The ausprivauto0405 dataset represents the Australian portfolio motor insurance market experience in 2004-2005.

There are 67.856 records with 9 variables; records in dataset represent all possible combinations of six risk classification variables: VehAge (the vehicle age, 4 levels), VehBody (the vehicle body, 13 levels), Gender (the gender of policyholder, 2 levels), DrivAge (the age of policyholder, 6 levels), VehValue (the vehicle value, 13 levels), ClaimOcc (indicates occurrence of claim). Each row contains the number of policies years (Exposure), number of claims (ClaimNb) and the sum of claim payments (ClaimAmount).

To create a commercial tariff [G.A. Spedicato] are defined some assumptions, first of all is assumed that ClaimAmount bundled payments, reserves and IBNR amounts, moreover there aren't projection about payment and frequency, so on commercial tariff is defined by:

$$\Pi^{\text{TAR}} = \frac{L_i + F}{(1 - H)}$$

$L_i$  denotes Loss Model, the model to represent risk premium;  $H=25\%$  represents premium fee subdivided into provisional fees and running fees, instead  $F = 20$  represents allowance for fixed costs.

In this work is considered risk premium at time  $t_0$  by the product between frequency model and severity model. This is a deterministic premium from which is determined the stochastic risk premium

necessary to evaluate new business and renewal portfolio premiums to create a business commercial tariff. In this work are elaborated two technical index: loss ratio and tariff requirement. The first one is a percentage of total losses incurred in claims respect total premiums earned, instead the second one is a balance constant used to cover actual losses with future earning premiums.

In this work are tested both univariate and multivariate analysis. In the first one are used categorized variables for GLM and GNM approaches obtained by cluster analysis, not for GAM approach because in this case cluster analysis doesn't found an exact correspondence with knots, so are used training tariff variables. For multivariate analysis are used covariates without transformation for computational reasons.

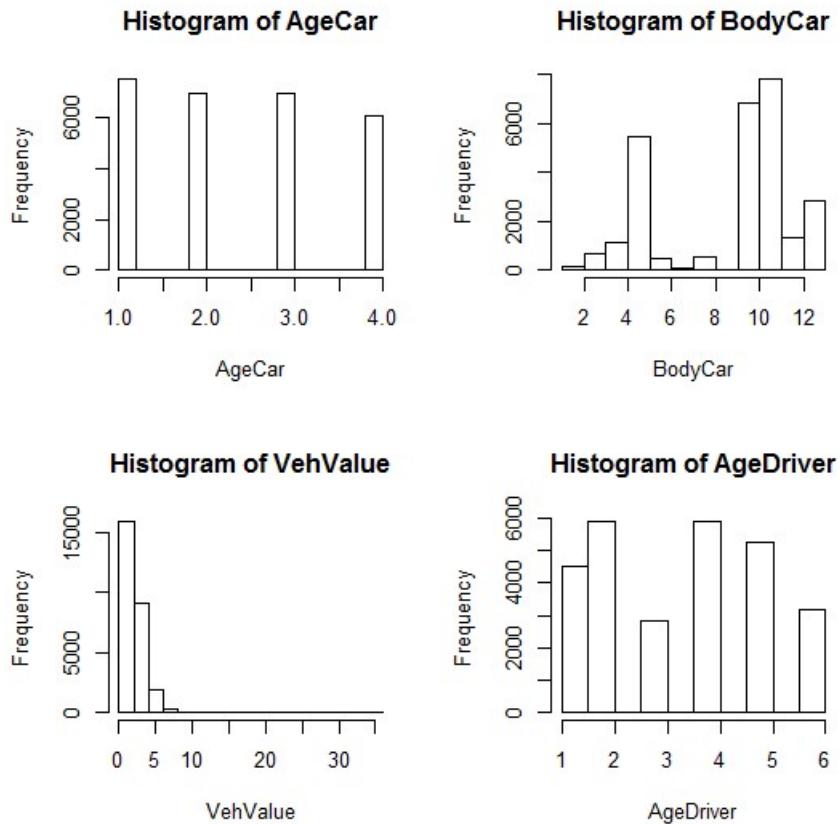
## Statistics

To work with this dataset first of all is necessary to organize data, starting to capture priority variables, (Gender and ClaimOcc are not relevant), as we could have more information about same policies on different rows, therefore is necessary to assemble it.

```
#data loading#
> library(CASdatasets)
> data(ausprivauto0405)
> attach(ausprivauto0405)
> motor<-aggregate(cbind(Exposure,ClaimNb,ClaimAmount)~VehValue
+ +ClaimOcc+DrivAge+VehAge+VehBody+Gender,na.action=na.omit,
+ data=ausprivauto0405, sum, order=T)
> motor$AgeDriver<-as.numeric(motor$DrivAge)
> motor$AgeCar<-as.numeric(motor$VehAge)
> motor$BodyCar<-as.numeric(motor$VehBody)
> rc<-motor[,c("AgeCar","BodyCar","VehValue","AgeDriver","Exposure",
+ "ClaimNb","ClaimAmount")]
```

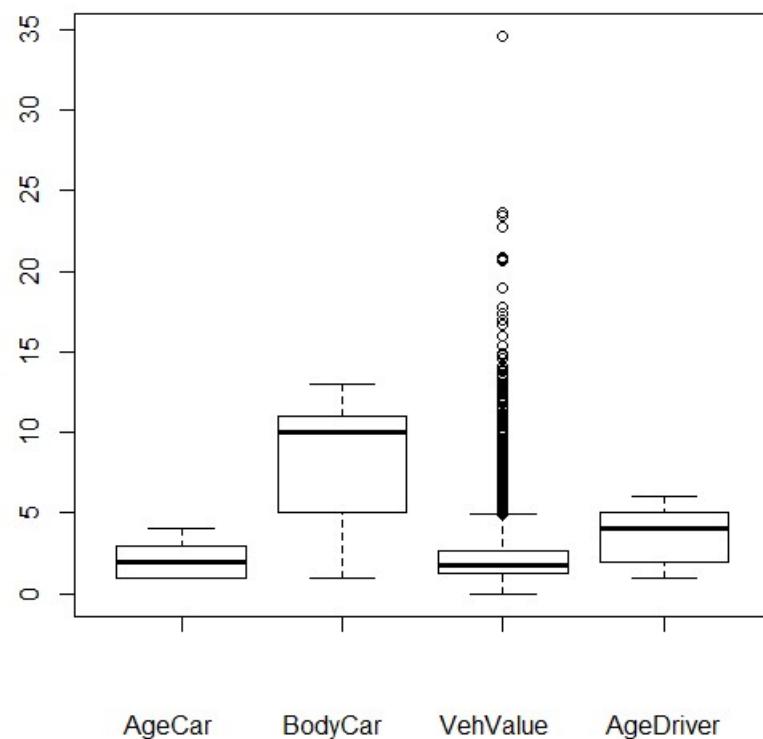
To understand what kind of approach is good to follow are used some descriptive graphic statistics such as histograms and boxplot.

```
#variables analysis#
> par(mfrow=c(2,2))#Pic.1 - Histograms of covariates#
> hist(rc$AgeCar, main="Histogram of AgeCar", xlab="AgeCar")
> hist(rc$BodyCar, main="Histogram of BodyCar", xlab="BodyCar")
> hist(rc$VehValue, main="Histogram of VehValue", xlab="VehValue")
> hist(rc$AgeDriver, main="Histogram of AgeDriver", xlab="AgeDriver")
> boxplot(rc$AgeCar,rc$BodyCar,rc$VehValue,rc$AgeDriver,
+ xlab="AgeCar BodyCar VehValue AgeDriver")#Pic.2 - boxplot#
```



Pic.1–Histograms of covariates

From this picture there aren't the same frequency distributions into each class of covariates.



Pic.2–Boxplot of covariates

This picture shows some outliers and likely they represent some withdrawal about amounts but they aren't removed.

```
#summaries statistics#
> summary(rc$AgeCar)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 1.000 2.000 2.418 3.000 4.000
> summary(rc$BodyCar)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 5.000 10.000 9.114 11.000 13.000
```

```

> summary(rc$VehValue)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  0.000 1.230 1.800 2.135 2.700 34.560

> summary(rc$AgeDriver)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  1.000 2.000 4.000 3.396 5.000 6.000

#mean frequency#
> MClaims<-with(rc, sum(ClaimNb)/sum(Exposure))
> MClaims
[1] 0.5471511

#mean severity#
> MACost<-with(rc, sum(ClaimAmount)/sum(ClaimNb))
> MACost
[1] 287.822

#mean risk premium#
> MPremium<-with(rc, sum(ClaimAmount)/sum(Exposure))
> MPremium
[1] 157.4821

```

Actuarial practice requires to work with variables split in tariff classes because the process about premium calculation needs to group variables into homogeneous risk classes, thus is used cluster analysis by k-means method that is faster than Ward method, but it requires to define number of clusters before.

The k-means clustering [Yu-Wei, 294-306] works partitioning  $n$  observations  $(x_1, x_2, \dots, x_n)$  into  $k$  clusters ( $k \leq n$ ) of the given set of data,  $S = (S_1, S_2, \dots, S_k)$  with the aim to minimize the within-cluster sum of squares (WCSS).

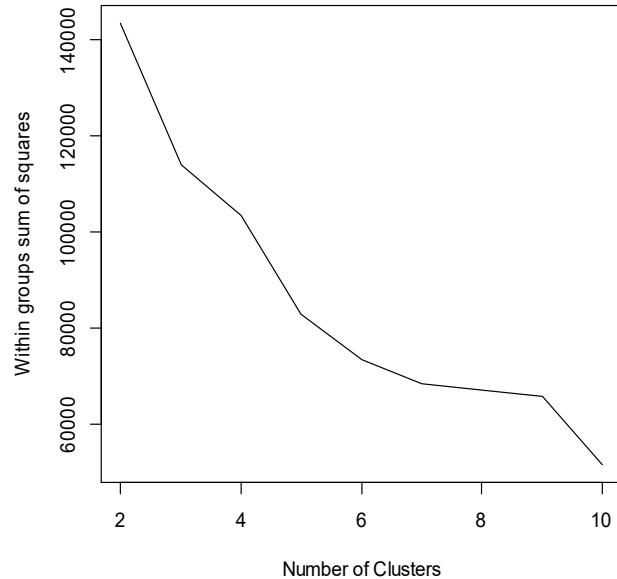
$$\arg \min_S \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

Where  $\mu_i$  is the mean of points of  $S_i$ .

First of all is needed to determine right number of clusters and one approach is by plot. The first

clusters add much information than the last ones where the marginal gain drop giving an angle in the graph; this is the point to choose right number of clusters.

```
#Prepare Data
> rc.stand<-scale(rc[-1]) # To standardize the variables
#Determine number of clusters
> nk = 2:10
> WSS = sapply(nk, function(k) {
+ kmeans(rc.stand, centers=k)$tot.withinss
+ })
> plot(nk, WSS, type="l", xlab="Number of Clusters",
+ ylab="Within groups sum of squares")
```



Pic.3-within cluster sum of squares plot

Looking the graph seven cluster appears right to classify risks into tariff variables.

```
#k-means with k = 7 solutions
> k.means.fit <- kmeans(rc.stand, 7)
```

```
#append cluster assignment
> rc.f<-data.frame(rc,k.means.fit$cluster)
#define risk classes into each variable by cut function
> rc.f$VehValue<-cut(rc$VehValue,7,right=F)
> rc.f$AgeDriver<-cut(rc$AgeDriver,7,right=F)
> rc.f$AgeCar<-cut(rc$AgeCar,7,right=F)
> rc.f$BodyCar<-cut(rc$BodyCar,7,right=F)
```

## Commercial MTPL Tariff with GLM Approach

Generalized Linear Models [E. Ohlsson, B. Johansson, pp. 15-35] are made up by a systematic component  $\eta_i = \sum_{j=1}^n x_{ij}\beta_j$  linked with a random component  $Y_i = EF(b(\theta_i); \frac{\phi_i}{\omega_i})$  by a link function  $g(\mu_i)$ .

Generalized Linear Models are an extension of linear models in two directions: by probability distribution with output variables that are stochastically independent with the same exponential family distribution and by expected value with a link function between mean value of outputs and covariates that could be different from linear regression.

$$E[Y|X] = \beta_0 + \sum_{j=1}^n x_{ij}\beta_j \Rightarrow g(E[Y|X]) = \beta_0 + \sum_{j=1}^n x_{ij}\beta_j$$

The structure of the variance is the same of the expected value:  $Var(Y_i) = \frac{\phi_i}{\omega_i} Var(\mu_i)$  with  $\mu_i = b'(\theta_i)$  and  $g(\mu_i) = b'^{-1}(\mu_i) = \theta_i$ .

## Univariate Analysis

For univariate analysis are used tariff variables grouped into homogeneous risk classes calculated before.

```
#univariate analysis with quasi Poisson and GLM approach for frequency model#
> ClaimsModglm1<-glm(ClaimNb~AgeCar+BodyCar+VehValue+AgeDriver
+ +offset(log(Exposure)), data=rc.f, family="quasipoisson"(link=log))
> anova(ClaimsModglm1, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F	Pr (>F)	
NULL			27520	13514			
AgeCar	3	122.45	27517	13391	77.4310	< 2.2e-16	***
BodyCar	6	432.51	27511	12958	136.7475	< 2.2e-16	***
VehValue	5	19.93	27506	12939	7.5599	4.185e-07	***
AgeDriver	5	92.56	27501	12846	35.1164	< 2.2e-16	***

All covariates are relevant for the frequency model by anova test.

```
> GLMClaims1<-predict(ClaimsModglm1,type="response")
> rc$GLMClaims1<-with(rc.f,predict(ClaimsModglm1,type="response"))
```

Claim severity needs the calculation about mean cost.

```
#univariate analysis with Normal and GLM approach for severity model#
> rc.f$ACost<-with(rc.f, ClaimAmount/ClaimNb)
> CostModglm1<-glm(ACost~AgeCar+BodyCar+VehValue+AgeDriver,
+ weights=ClaimNb, data=rc.f,family="gaussian"(link=log))
> anova(CostModglm1, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F	Pr (>F)	
NULL			27520	3.2697e+10			
AgeCar	3	34869267	27517	3.2662e+10	9.8482	1.733e-06	***
BodyCar	6	110260625	27511	3.2552e+10	15.5706	< 2.2e-16	***
VehValue	5	6902713	27506	3.2545e+10	1.1697	0.3213	
AgeDriver	5	88808479	27501	3.2456e+10	15.0495	8.661e-15	***

Not all covariates are relevant by analysis of variance so, with Backward method, VehValue variable is eliminated by the model with new testing. Also in the severity model is used a Gaussian distribution because with a Gamma distribution the model doesn't converge.

```
> CostModglm1<-update(CostModglm1,~.-VehValue)
> anova(CostModglm1, test="F")

          Df Deviance Resid. Df Resid. Dev F Pr(>F)
NULL             27520 3.2697e+10
AgeCar          3 34869267 27517 3.2662e+10 9.8469 1.737e-06 ***
BodyCar         6 110260625 27511 3.2552e+10 15.5685 < 2.2e-16 ***
AgeDriver       5 89063932 27506 3.2463e+10 15.0907 7.846e-15 ***
```

Now all covariates selected are relevant.

```
> GLMCost1<-predict(CostModglm1,type="response")
> rc$GLMCost1<-with(rc.f,predict(CostModglm1,type="response"))
```

After that is calculated deterministic risk premium and so on stochastic risk premium.

```
#deterministic risk premium#
> rc.f$RiskPremium1<-with(rc, (GLMCost1*GLMClaims1)/Exposure)

#stochastic risk premium with GLM approach#
> PRSModglm1<-glm(RiskPremium1~AgeCar+BodyCar+VehValue+AgeDriver,
+ weights=Exposure, data=rc.f, family=gaussian(link=log))
> GLMSRiskPremium1<-predict(PRSModglm1,data=rc.f,type="response")
> rc$GLMSRiskPremium1<-with(rc.f, GLMSRiskPremium1)
```

Index for the goodness of the model is done by the difference between expected losses and actual losses.

```
#Expected losses and actual losses
> expectedlosses<-with(rc.f, sum(GLMSRiskPremium1*Exposure))
```

```

> expectedlosses
[1] 9337547
> actuallosses<-with(rc.f, sum(ClaimAmount) )
> actuallosses
[1] 9342125
> round((expectedlosses-actuallosses)/actuallosses,4)#difference#
[1] -5e-04

```

With positive result next step is to apply loadings and tariff requirement at the risk premium.

```

#fees#
> F<-20
> H<-0.25

#tariff requirement#
> GLMFabTariff1<-with(rc.f,
+ sum(ClaimAmount) /sum(GLMSRiskPremium1*Exposure) )
> GLMFabTariff1
[1] 1.00049

#Commercial tariff#
> rc.f$GLMTariffPremium1<-with(rc.f,
+ (GLMSRiskPremium1*GLMFabTariff1+F) / (1-H) )

#Mean commercial tariff#
> MGLMTariffPremium1<-with(rc.f, mean(GLMTariffPremium1))
> MGLMTariffPremium1
[1] 234.4587

```

```
#Loss ratio#
> with(rc.f, sum(ClaimAmount) / sum(GLMTariffPremium1))
[1] 1.447822
```

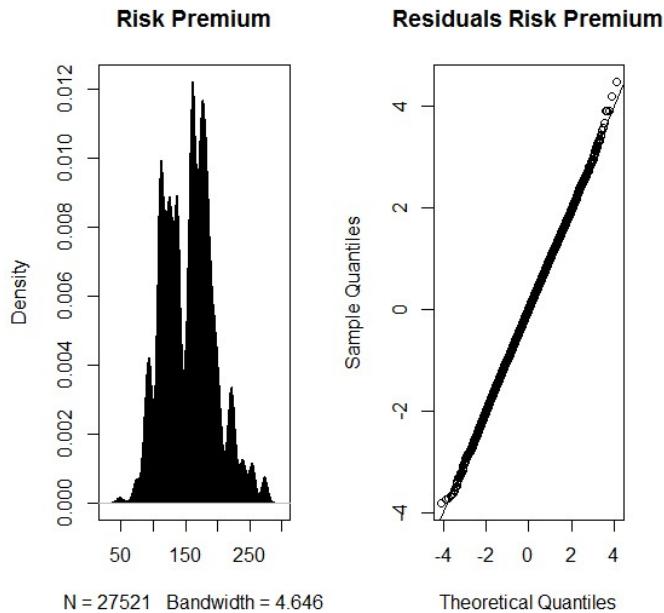
There are twenty risk coefficients with the evidence that selecting AgeDriver variable in the last class premium is double of the intercept.

```
#Tariff coefficients#
> coefRiskPremium1<- (coef(PRSModglm1))
> tableCoeff<-data.frame(rating.factor=
+ c(levels<-names(coefRiskPremium1)),
+ coefRiskPremium=exp(as.numeric(coefRiskPremium1)))
> print(tableCoeff,digits=2)
```

	rating.factor	coefRiskPremium
1	(Intercept)	73.08
2	AgeCar[1.86,2.29)	1.01
3	AgeCar[2.71,3.14)	1.08
4	AgeCar[3.57,4)	0.85
5	BodyCar[2.71,4.43)	1.58
6	BodyCar[4.43,6.14)	1.73
7	BodyCar[6.14,7.86)	0.65
8	BodyCar[7.86,9.57)	1.19
9	BodyCar[9.57,11.3)	1.52
10	BodyCar[11.3,13)	1.19
11	VehValue[4.94,9.87)	1.15
12	VehValue[9.87,14.8)	1.15
13	VehValue[14.8,19.7)	1.02
14	VehValue[19.7,24.7)	1.09
15	VehValue[29.6,34.6)	0.72
16	AgeDriver[1.71,2.43)	1.45

17	AgeDriver[2.43, 3.14)	1.01
18	AgeDriver[3.86, 4.57)	1.45
19	AgeDriver[4.57, 5.29)	1.61
20	AgeDriver[5.29, 6)	2.00

```
#Pic.4 - Distributions Risk Premium and residuals#
> par(mfrow=c(1,2))
> b<-density(GLMSRiskPremium1)
> plot(b, type="h", col="black", main="Risk Premium")
> qqnorm(rnorm(resid(PRSModglm1)), main="Residuals Risk Premium")
> qqline(rnorm(resid(PRSModglm1)))
```



Pic.4 - Distributions Risk Premium and residuals for univariate GLM

```
#test on the model#
> summary(PRSModglm1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
7.800e-13	7.478e-12	9.494e-12	1.168e-11	4.418e-11

Coefficients:

	Estimate	Std. Error	t value	Pr (> t )
(Intercept)	4.292e+00	6.941e-15	6.183e+14	$<2e-16$ ***
AgeCar[1.86, 2.29)	1.087e-02	7.350e-16	1.478e+13	$<2e-16$ ***
AgeCar[2.71, 3.14)	7.590e-02	7.139e-16	1.063e+14	$<2e-16$ ***
AgeCar[3.57, 4)	-1.590e-01	8.976e-16	-1.772e+14	$<2e-16$ ***
BodyCar[2.71, 4.43)	4.560e-01	6.958e-15	6.553e+13	$<2e-16$ ***
BodyCar[4.43, 6.14)	5.459e-01	6.888e-15	7.926e+13	$<2e-16$ ***
BodyCar[6.14, 7.86)	-4.279e-01	1.437e-14	-2.978e+13	$<2e-16$ ***
BodyCar[7.86, 9.57)	1.743e-01	7.336e-15	2.376e+13	$<2e-16$ ***
BodyCar[9.57, 11.3)	4.181e-01	6.873e-15	6.084e+13	$<2e-16$ ***
BodyCar[11.3, 13)	1.738e-01	6.939e-15	2.505e+13	$<2e-16$ ***
VehValue[4.94, 9.87)	1.408e-01	1.653e-15	8.513e+13	$<2e-16$ ***
VehValue[9.87, 14.8)	1.403e-01	7.298e-15	1.922e+13	$<2e-16$ ***
VehValue[14.8, 19.7)	2.155e-02	2.118e-14	1.017e+12	$<2e-16$ ***
VehValue[19.7, 24.7)	8.735e-02	2.373e-14	3.682e+12	$<2e-16$ ***
VehValue[29.6, 34.6)	-3.304e-01	1.234e-13	-2.678e+12	$<2e-16$ ***
AgeDriver[1.71, 2.43)	3.726e-01	1.128e-15	3.303e+14	$<2e-16$ ***
AgeDriver[2.43, 3.14)	6.134e-03	1.561e-15	3.929e+12	$<2e-16$ ***
AgeDriver[3.86, 4.57)	3.689e-01	1.132e-15	3.260e+14	$<2e-16$ ***
AgeDriver[4.57, 5.29)	4.757e-01	1.128e-15	4.218e+14	$<2e-16$ ***
AgeDriver[5.29, 6)	6.947e-01	1.166e-15	5.956e+14	$<2e-16$ ***

(Dispersion parameter for gaussian family taken to be 1.154268e-22)

```
Null deviance: 8.8947e+07 on 27520 degrees of freedom
Residual deviance: 3.1744e-18 on 27501 degrees of freedom
```

```
AIC: -1329485
```

```
Number of Fisher Scoring iterations: 1
```

```
> R2glm1<-cor(rc.f$RiskPremium1,predict(PRSModglm1, type="link"))^2
> R2glm1
[1] 0.9696606
> AIC(PRSModglm1)
[1] -1329485
> BIC(PRSModglm1)
[1] -1329312
> deviance(PRSModglm1)
[1] 3.174353e-18
> df.residual(PRSModglm1)
[1] 27501
```

Deviance explained by this model is very high.

## Multivariate Analysis

With multivariate approach is possible to investigate about correlation between variables, so this methodology can be considered an upgrade respect the previous one.

```
#multivariate analysis with quasi Poisson and GLM approach for frequency
model#
> ClaimsModglm2<-glm(ClaimNb~AgeCar*BodyCar*VehValue*AgeDriver
+ +offset(log(Exposure)),data=rc,family="quasipoisson"(link=log))
> anova(ClaimsModglm2, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F
NULL			27520	13514	
AgeCar	1	95.303	27519	13418 181.5661	
BodyCar	1	51.380	27518	13367 97.8872	

VehValue	1	237.634	27517	13129	452.7291
AgeDriver	1	57.326	27516	13072	109.2147
AgeCar:BodyCar	1	7.239	27515	13065	13.7917
AgeCar:VehValue	1	56.299	27514	13008	107.2589
BodyCar:VehValue	1	5.392	27513	13003	10.2729
AgeCar:AgeDriver	1	0.434	27512	13002	0.8272
BodyCar:AgeDriver	1	1.560	27511	13001	2.9712
VehValue:AgeDriver	1	7.587	27510	12993	14.4539
AgeCar:BodyCar:VehValue	1	0.536	27509	12993	1.0210
AgeCar:BodyCar:AgeDriver	1	0.161	27508	12993	0.3076
AgeCar:VehValue:AgeDriver	1	0.013	27507	12993	0.0243
BodyCar:VehValue:AgeDriver	1	0.395	27506	12992	0.7533
AgeCar:BodyCar:VehValue:AgeDriver	1	0.950	27505	12991	1.8095

Pr (>F)

NULL

AgeCar	< 2.2e-16	***
BodyCar	< 2.2e-16	***
VehValue	< 2.2e-16	***
AgeDriver	< 2.2e-16	***
AgeCar:BodyCar	0.0002046	***
AgeCar:VehValue	< 2.2e-16	***
BodyCar:VehValue	0.0013515	**
AgeCar:AgeDriver	0.3630996	
BodyCar:AgeDriver	0.0847711	.
VehValue:AgeDriver	0.0001439	***
AgeCar:BodyCar:VehValue	0.3122866	
AgeCar:BodyCar:AgeDriver	0.5791524	
AgeCar:VehValue:AgeDriver	0.8760699	
BodyCar:VehValue:AgeDriver	0.3854351	
AgeCar:BodyCar:VehValue:AgeDriver	0.1785773	

From analysis of variance some parameters are useless so are removed.

```
> ClaimsModglm2<-update(ClaimsModglm2, ~.-AgeCar:AgeDriver
+ -AgeCar:BodyCar:VehValue-AgeCar:BodyCar:AgeDriver
+ -AgeCar:VehValue:AgeDriver
+ -BodyCar:VehValue:AgeDriver-AgeCar:BodyCar:VehValue:AgeDriver)
> anova(ClaimsModglm2, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F	Pr (>F)
NULL		27520	13514			
AgeCar	1	95.303	27519	13418 181.6091	< 2.2e-16	***
BodyCar	1	51.380	27518	13367 97.9104	< 2.2e-16	***
VehValue	1	237.634	27517	13129 452.8363	< 2.2e-16	***
AgeDriver	1	57.326	27516	13072 109.2406	< 2.2e-16	***
AgeCar:BodyCar	1	7.239	27515	13065 13.7950	0.0002043	***
AgeCar:VehValue	1	56.299	27514	13008 107.2843	< 2.2e-16	***
BodyCar:VehValue	1	5.392	27513	13003 10.2754	0.0013497	**
BodyCar:AgeDriver	1	1.581	27512	13001 3.0124	0.0826429	.
VehValue:AgeDriver	1	7.730	27511	12994 14.7310	0.0001243	***

In this way are captured all relevant variables included those correlated.

```
> GLMClaims2<-predict(ClaimsModglm2, type="response")
> rc$GLMClaims2<-with(rc, predict(ClaimsModglm2, type="response"))

#multivariate analysis with Normal and GLM approach for severity model#
> rc$ACost<-with(rc, ClaimAmount/ClaimNb)
> CostModglm2<-glm(ACost~AgeCar*BodyCar*VehValue*AgeDriver,
+ weights=ClaimNb, data=rc, family=gaussian(link=log))
> anova(CostModglm2, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F
NULL			27520	3.2697e+10	
AgeCar	1	19396254	27519	3.2678e+10	16.4128
BodyCar	1	35198521	27518	3.2643e+10	29.7843
VehValue	1	45374901	27517	3.2597e+10	38.3954
AgeDriver	1	33145985	27516	3.2564e+10	28.0475
AgeCar:BodyCar	1	1884260	27515	3.2562e+10	1.5944
AgeCar:VehValue	1	385686	27514	3.2562e+10	0.3264
BodyCar:VehValue	1	9896121	27513	3.2552e+10	8.3739
AgeCar:AgeDriver	1	3237929	27512	3.2549e+10	2.7399
BodyCar:AgeDriver	1	246769	27511	3.2548e+10	0.2088
VehValue:AgeDriver	1	35705123	27510	3.2513e+10	30.2130
AgeCar:BodyCar:VehValue	1	2038901	27509	3.2511e+10	1.7253
AgeCar:BodyCar:AgeDriver	1	38430	27508	3.2511e+10	0.0325
AgeCar:VehValue:AgeDriver	1	93951	27507	3.2510e+10	0.7950
BodyCar:VehValue:AgeDriver	1	570696	27506	3.2509e+10	0.4829
AgeCar:BodyCar:VehValue:AgeDriver	1	692985	27505	3.2508e+10	0.5864
				Pr (>F)	
NULL					
AgeCar		5.108e-05		***	
BodyCar		4.870e-08		***	
VehValue		5.859e-10		***	
AgeDriver		1.193e-07		***	
AgeCar:BodyCar		0.206706			
AgeCar:VehValue		0.567815			
BodyCar:VehValue		0.003809		**	
AgeCar:AgeDriver		0.097884		.	
BodyCar:AgeDriver		0.647704			
VehValue:AgeDriver		3.905e-08		***	
AgeCar:BodyCar:VehValue		0.189026			
AgeCar:BodyCar:AgeDriver		0.856894			

```

AgeCar:VehValue:AgeDriver          0.372601
BodyCar:VehValue:AgeDriver         0.487112
AgeCar:BodyCar:VehValue:AgeDriver  0.443825

> CostModglm2<-update(CostModglm2,~.-AgeCar:BodyCar-AgeCar:VehValue
+ -BodyCar:AgeDriver-AgeCar:BodyCar:VehValue-AgeCar:BodyCar:AgeDriver
+ -AgeCar:VehValue:AgeDriver-BodyCar:VehValue:AgeDriver
+ -AgeCar:BodyCar:VehValue:AgeDriver)
> anova(CostModglm2, test="F")

      Df Deviance Resid.Df Resid.Dev      F      Pr (>F)
NULL           27520 3.2697e+10
AgeCar          1 19396254  27519 3.2678e+10 16.4071 5.123e-05 ***
BodyCar          1 35198521  27518 3.2643e+10 29.7740 4.896e-08 ***
VehValue         1 45374901  27517 3.2597e+10 38.3821 5.899e-10 ***
AgeDriver        1 33145985  27516 3.2564e+10 28.0378 1.199e-07 ***
BodyCar:VehValue 1 5527080  27515 3.2558e+10  4.6753  0.03061   *
AgeCar:AgeDriver 1 3956829  27514 3.2555e+10  3.3470  0.06734   .
VehValue:AgeDriver 1 31592860  27513 3.2523e+10 26.7240 2.363e-07 ***

```

In the severity model such as in the frequency model univariate variables are significant and only some multivariate ones.

```

> GLMCost2<-predict(CostModglm2,type="response")
> rc$GLMCost2<-with(rc,predict(CostModglm2,type="response"))

#deterministic risk premium#
> rc$RiskPremium2<-with(rc, (GLMCost2*GLMClaims2)/Exposure)

#stochastic risk premium with GLM approach#
> PRSModglm2<-glm(RiskPremium2~AgeCar*BodyCar*VehValue*AgeDriver,

```

```
+ weights=Exposure, data=rc, family=gaussian(link=log))
> GLMSRiskPremium2<-predict(PRSModglm2,data=rc,type="response")

#Expected losses and actual losses
> expectedlosses<-with(rc, sum(GLMSRiskPremium2*Exposure))
> expectedlosses
[1] 9359678
> actuallosses<-with(rc, sum(ClaimAmount))
> actuallosses
[1] 9342125
> round((expectedlosses-actuallosses)/actuallosses,4) #difference#
[1] 0.0019
```

Multivariate analysis generates an overestimation of the expected losses and so on tariff requirement is lower than univariate analysis.

```
#fees#
> F<-20
> H<-0.25

#tariff requirement#
> GLMFabTariff2<-with(rc,sum(ClaimAmount) /
+ sum(GLMSRiskPremium2*Exposure))
> GLMFabTariff2
[1] 0.9981246

#Commercial tariff#
> rc$GLMTariffPremium2<-with(rc,
+ (GLMSRiskPremium2*GLMFabTariff2+F) / (1-H))
```

```
#Mean commercial tariff#
> MGLMTariffPremium2<-with(rc, mean(GLMTariffPremium2))
> MGLMTariffPremium2
[1] 234.6486

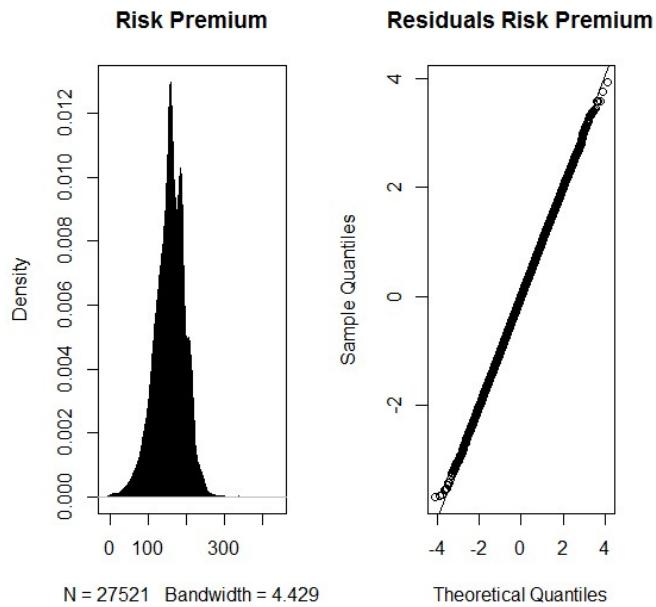
#Loss ratio#
> with(rc, sum(ClaimAmount) / sum(GLMTariffPremium2))
[1] 1.44665

#Tariff coefficients#
> coefRiskPremium2<-(coef(PRSModglm2))
> tableCoeff<-data.frame(rating.factor=
+ c(levels<-names(coefRiskPremium2)),
+ coefRiskPremium=exp(as.numeric(coefRiskPremium2)))
> print(tableCoeff,digits=2)
      rating.factor  coefRiskPremium
1             (Intercept)     251.03
2                 AgeCar       1.00
3                 BodyCar       0.95
4                 VehValue       0.72
5                 AgeDriver      1.01
6             AgeCar:BodyCar     1.01
7             AgeCar:VehValue     0.97
8             BodyCar:VehValue     1.01
9             AgeCar:AgeDriver     1.00
10            BodyCar:AgeDriver     1.00
11            VehValue:AgeDriver     1.06
12            AgeCar:BodyCar:VehValue     1.00
13            AgeCar:BodyCar:AgeDriver     1.00
14            AgeCar:VehValue:AgeDriver     1.00
15            BodyCar:VehValue:AgeDriver     1.00
```

16      AgeCar:BodyCar:VehValue:AgeDriver	1.00
---	------

With multivariate approach collective mean tariff premium is quite the same of the univariate approach, but there are less risk coefficients with lower value respect the previous before. The overestimation is reflected on the increased intercept.

```
#Distributions Risk Premium and residuals#
> par(mfrow=c(1,2))
> b<-density(GLMSRiskPremium2)
> plot(b,type="h", col="black",main="Risk Premium")
> x<-rnorm(resid(PRSModglm2))
> qqnorm(x,main="Residuals Risk Premium")
> qqline(x)
```



Pic.5 - Distributions Risk Premium and residuals for multivariate GLM

```
#test on the model#
> summary(PRSModglm2)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.500e-13	1.800e-12	2.544e-12	3.467e-12	3.342e-11

Coefficients:

	Estimate	Std. Error	t value	Pr (> t )
(Intercept)	5.526e+00	3.509e-15	1.575e+15	<2e-16***
AgeCar	1.187e-03	1.561e-15	7.604e+11	<2e-16***
BodyCar	-5.552e-02	3.927e-16	-1.414e+14	<2e-16***
VehValue	-3.307e-01	2.314e-15	-1.429e+14	<2e-16***
AgeDriver	1.232e-02	7.903e-16	1.558e+13	<2e-16***
AgeCar:BodyCar	9.173e-03	1.718e-16	5.340e+13	<2e-16***
AgeCar:VehValue	-2.852e-02	9.140e-16	-3.120e+13	<2e-16***
BodyCar:VehValue	1.155e-02	2.369e-16	4.878e+13	<2e-16***
AgeCar:AgeDriver	-4.754e-03	3.391e-16	-1.402e+13	<2e-16***
BodyCar:AgeDriver	-6.870e-04	8.928e-17	-7.695e+12	<2e-16***
VehValue:AgeDriver	6.130e-02	5.024e-16	1.220e+14	<2e-16***
AgeCar:BodyCar:VehValue	-5.404e-16	9.125e-17	-5.922e+00	3.22e-09***
AgeCar:BodyCar:AgeDriver	-6.164e-17	3.766e-17	-1.637e+00	0.102
AgeCar:VehValue:AgeDriver	-1.152e-15	1.912e-16	-6.024e+00	1.72e-09***
BodyCar:VehValue:AgeDriver	-3.637e-16	5.169e-17	-7.036e+00	2.02e-12***
AgeCar:BodyCar:VehValue:AgeDriver	1.038e-16	1.923e-17	5.396e+00	6.89e-08***

(Dispersion parameter for gaussian family taken to be 9.4853e-24)

Null deviance: 7.6299e+07 on 27520 degrees of freedom  
 Residual deviance: 2.6089e-19 on 27505 degrees of freedom  
 AIC: -1398261

Number of Fisher Scoring iterations: 1

```
> R2glm2<-cor(rc$RiskPremium2,predict(PRSModglm2, type="link"))^2
```

```

> R2glm2
[1] 0.8763893
> AIC(PRSModglm2)
[1] -1398261
> BIC(PRSModglm2)
[1] -1398121
> deviance(PRSModglm2)
[1] 2.608932e-19
> df.residual(PRSModglm2)
[1] 27505

```

Analyzing results last method shows less risk coefficients and less deviance explained (87,64%) respect univariate analysis (96,96%). This good result from univariate approach is also confirmed comparing it with the deviance explained (98,41%) obtained by training variables with an error rate about of -1,47%.

## Commercial MTPL Tariff with GNM Approach

A Generalized Non-linear Model [H. Turner, D. Firth] is the same as the GLM model except link function  $g(\mu_i) = \eta_i(x_{ij}; \beta_j)$  where the systematic component is non-linear in the parameters  $\beta_j$ . GNM may also be considered as an extension of a least squares model where the variance of the output depends from the expected value. This model works both with exponential models and with multiplicative models. This model requires an iterative method to estimate non-linear parameters because in this situation likelihood may have multiple optimal solutions and also, such as GLM, it can't handle random variables in the predictor.

## Univariate Analysis

Generalized Non linear approach requires to upload gnm library, after that the univariate analysis follows the same path used previously in glm. About frequency model is used quasi Poisson distribution for all statistical models shows in this paper to avoid to run overdispersion.

```

> library(gnm)
#univariate analysis with quasi Poisson and GNM approach#
> ClaimsModgnm1<-gnm(ClaimNb~AgeCar+BodyCar+VehValue+AgeDriver
+ +offset(log(Exposure)),data=rc.f,family="quasipoisson"(link=log))
> anova(ClaimsModgnm1, test="F")

```

	Df	Deviance	Resid.Df	Resid.Dev	F	Pr (>F)
NULL			27520	13514		
AgeCar	3	122.45	27517	13391	77.3998	< 2.2e-16 ***
BodyCar	6	432.51	27511	12958	136.6924	< 2.2e-16 ***
VehValue	5	19.93	27506	12939	7.5569	4.215e-07 ***
AgeDriver	5	92.56	27501	12846	35.1023	< 2.2e-16 ***

All covariates are relevant for Claims model, not the same for Cost model, so is used Backward test for the last one model until to have all relevant independent variables inside it. Also is used Gaussian distribution such as GLM because model doesn't converge.

```

> GNMClaims1<-predict(ClaimsModgnm1,type="response")
> rc$GNMClaims1<-with(rc.f,predict(ClaimsModgnm1,type="response"))

```

```

#univariate analysis with Gaussian and GNM approach#
> rc.f$ACost<-with(rc.f, ClaimAmount/ClaimNb)
> CostModgnm1<-gnm(ACost~AgeCar+BodyCar+VehValue+AgeDriver,
+ weights=ClaimNb, data=rc.f,family="gaussian"(link=log))
> anova(CostModgnm1, test="F")

```

	Df	Deviance	Resid.Df	Resid.Dev	F	Pr (>F)
NULL			27520	3.2697e+10		
AgeCar	3	34869267	27517	3.2662e+10	9.8218	1.801e-06 ***
BodyCar	6	110260625	27511	3.2552e+10	15.5289	< 2.2e-16 ***
VehValue	5	6902713	27506	3.2545e+10	1.1666	0.3228

```
AgeDriver 5 88807878 27501 3.2456e+10 15.0090 9.543e-15 ***
```

```
> CostModgnm1<-update(CostModgnm1, ~.-VehValue)
> anova(CostModgnm1, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F	Pr(>F)
NULL			27520	3.2697e+10		
AgeCar	3	34869267	27517	3.2662e+10	9.8283	1.784e-06 ***
BodyCar	6	110260625	27511	3.2552e+10	15.5391	< 2.2e-16 ***
AgeDriver	5	89063700	27506	3.2463e+10	15.0622	8.402e-15 ***

```
> GNMCost1<-predict(CostModgnm1, type="response")
> rc$GNMCost1<-with(rc.f, predict(CostModgnm1, type="response"))
```

For stochastic risk premium GNM is able to run with Gamma distribution.

```
#deterministic risk premium#
> rc.f$RiskPremium1<-with(rc, (GNMCost1*GNMClaims1)/Exposure)
```

```
#stochastic risk premium with GNM approach#
> PRSModgnm1<-gnm(RiskPremium1~AgeCar+BodyCar+VehValue
+ +AgeDriver, weights=Exposure, data=rc.f, family=Gamma(link=log))
> GNMSRiskPremium1<-predict(PRSModgnm1, data=rc.f, type="response")
> rc.f$GNMSRiskPremium1<-with(rc.f, GNMSRiskPremium1)
```

Results about expected losses are slightly better than GLM and so on tariff requirement, mean collective premium and Loss Ratio.

```
#Expected losses and actual losses
> expectedlosses<-with(rc.f, sum(GNMSRiskPremium1*Exposure))
```

```
> expectedlosses
[1] 9337683
> actuallosses<-with(rc.f, sum(ClaimAmount))
> actuallosses
[1] 9342125
> round((expectedlosses-actuallosses)/actuallosses,4)#difference#
[1] -5e-04

#fees#
> F<-20
> H<-0.25

#tariff requirement#
> GNMFabTariff1<-with(rc.f,sum(ClaimAmount) /
+ sum(GNMSRiskPremium1*Exposure))
> GNMFabTariff1
[1] 1.000476

#Commercial tariff#
> rc.f$GNMTariffPremium1<-with(rc.f,
+ (GNMSRiskPremium1*GNMFabTariff1+F) / (1-H))

#Mean commercial tariff#
> MGNMTariffPremium1<-with(rc.f, mean(GNMTariffPremium1))
> MGNMTariffPremium1
[1] 234.4647

#Loss ratio#
> with(rc.f, sum(ClaimAmount) / sum(GNMTariffPremium1))
[1] 1.447785
```

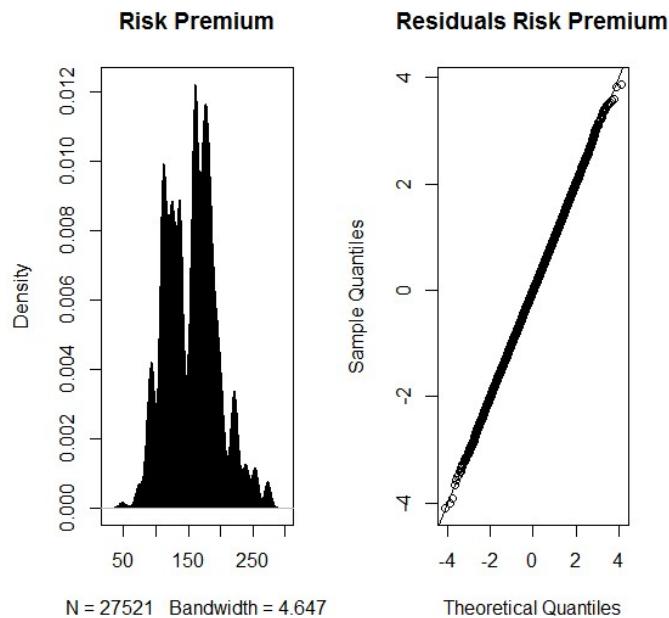
```
#Tariff coefficients#
```

```
> coefRiskPremium1<- (coef(PRSModgnm1))
> tableCoeff<-data.frame(coefficient=
+ c(levels<-names(coefRiskPremium1)),
+ coefficientsRiskPremium=exp(as.numeric(coefRiskPremium1)))
> print(tableCoeff,digits=2)
```

	coefficient	coefficientsRiskPremium
1	(Intercept)	73.01
2	AgeCar[1.86,2.29)	1.01
3	AgeCar[2.71,3.14)	1.08
4	AgeCar[3.57,4)	0.85
5	BodyCar[2.71,4.43)	1.58
6	BodyCar[4.43,6.14)	1.73
7	BodyCar[6.14,7.86)	0.65
8	BodyCar[7.86,9.57)	1.19
9	BodyCar[9.57,11.3)	1.52
10	BodyCar[11.3,13)	1.19
11	VehValue[4.94,9.87)	1.15
12	VehValue[9.87,14.8)	1.15
13	VehValue[14.8,19.7)	1.02
14	VehValue[19.7,24.7)	1.09
15	VehValue[29.6,34.6)	0.72
16	AgeDriver[1.71,2.43)	1.45
17	AgeDriver[2.43,3.14)	1.01
18	AgeDriver[3.86,4.57)	1.45
19	AgeDriver[4.57,5.29)	1.61
20	AgeDriver[5.29,6)	2.00

Risk coefficients are the same of GLM model, only the intercept is slightly better, also the shape of the risk premium distribution is equal of the previously model.

```
#Distributions Risk Premium and residuals#
> par(mfrow=c(1,2))
> b<-density(GNMSRiskPremium1)
> plot(b,type="h", col="black",main="Risk Premium")
> x<-rnorm(resid(PRSModgnm1))
> qqnorm(x,main="Residuals Risk Premium")
> qqline(x)
```



Pic.6 - Distributions Risk Premium and residuals for univariate GNM

```
#test on the model#
> summary(PRSModgnm1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.293e-08	-1.002e-08	-2.268e-09	0.000e+00	0.000e+00

Coefficients:

	Estimate	Std.Error	t value	Pr(> t )
(Intercept)	4.291e+00	0.000e+00	1.519e+15	<2e-16 ***
AgeCar[1.86, 2.29)	1.063e-02	0.000e+00	2.600e+13	<2e-16 ***
AgeCar[2.71, 3.14)	7.586e-02	0.000e+00	1.818e+14	<2e-16 ***
AgeCar[3.57, 4)	-1.587e-01	0.000e+00	-3.448e+14	<2e-16 ***
BodyCar[2.71, 4.43)	4.569e-01	0.000e+00	1.589e+14	<2e-16 ***
BodyCar[4.43, 6.14)	5.467e-01	0.000e+00	1.941e+14	<2e-16 ***
BodyCar[6.14, 7.86)	-4.265e-01	0.000e+00	-1.081e+14	<2e-16 ***
BodyCar[7.86, 9.57)	1.741e-01	0.000e+00	5.748e+13	<2e-16 ***
BodyCar[9.57, 11.3)	4.187e-01	0.000e+00	1.495e+14	<2e-16 ***
BodyCar[11.3, 13)	1.748e-01	0.000e+00	6.173e+13	<2e-16 ***
VehValue[4.94, 9.87)	1.408e-01	0.000e+00	1.481e+14	<2e-16 ***
VehValue[9.87, 14.8)	1.403e-01	0.000e+00	3.536e+13	<2e-16 ***
VehValue[14.8, 19.7)	2.155e-02	0.000e+00	2.303e+12	<2e-16 ***
VehValue[19.7, 24.7)	8.735e-02	0.000e+00	6.718e+12	<2e-16 ***
VehValue[29.6, 34.6)	-3.304e-01	2.768e-14	-1.193e+13	<2e-16 ***
AgeDriver[1.71, 2.43)	3.728e-01	0.000e+00	7.549e+14	<2e-16 ***
AgeDriver[2.43, 3.14)	6.606e-03	0.000e+00	1.077e+13	<2e-16 ***
AgeDriver[3.86, 4.57)	3.691e-01	0.000e+00	7.452e+14	<2e-16 ***
AgeDriver[4.57, 5.29)	4.760e-01	0.000e+00	9.226e+14	<2e-16 ***
AgeDriver[5.29, 6)	6.952e-01	0.000e+00	1.124e+15	<2e-16 ***

(Dispersion parameter for Gamma family taken to be 1.374923e-27)

Residual deviance: 1.3855e-14 on 27501 degrees of freedom

AIC: -1839415

Number of iterations: 1

```
> R2gnm1<-cor(rc.f$RiskPremium1,predict(PRSModgnm1, type="link"))^2
```

```

> R2gnm1
[1] 0.9696561
> BIC(PRSModgnm1)
[1] -1839242
> AIC(PRSModgnm1)
[1] -1839415
> deviance(PRSModgnm1)
[1] 1.385516e-14
> df.residual(PRSModgnm1)
[1] 27501

```

GNM results are quite the same of GLM model because variables have been run in linear way so the expectation are confirmed. Looking anova test, GNM model is more confident because residual deviance is lower than the previous one.

```

> anova(PRSModglm1, PRSModgnm1, test="F")
Analysis of Deviance Table

Model 1: RiskPremium1 ~ AgeCar + BodyCar + VehValue + AgeDriver
Model 2: RiskPremium1 ~ AgeCar + BodyCar + VehValue + AgeDriver

  Resid.Df  Resid.Dev Df   Deviance F Pr(>F)
1      27501 3.2000e-18
2      27501 1.3855e-14  0  -1.3852e-14

```

## Multivariate Analysis

Multivariate approach follows the previously GLM approach with Backward method to determine the main variables for the estimate model.

```

#multivariate analysis with quasi Poisson and GNM approach#
> ClaimsModgnm2<-gnm(ClaimNb~VehValue*AgeDriver*AgeCar*BodyCar
+ +offset(log(Exposure)), data=rc, family="quasipoisson"
+ (link=log))

```

```
> anova(ClaimsModgnm2, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F
NULL		27520	13514		
VehValue	1	366.42	27519	13147	697.8739
AgeDriver	1	56.95	27518	13090	108.4687
VehValue:AgeDriver	1	11.98	27517	13078	22.8238
AgeCar	1	6.95	27516	13071	13.2313
VehValue:AgeCar	1	44.61	27515	13027	84.9727
AgeDriver:AgeCar	1	0.21	27514	13026	0.4061
VehValue:AgeDriver:AgeCar	1	0.12	27513	13026	0.2248
BodyCar	1	9.40	27512	13017	17.9116
VehValue:BodyCar	1	0.56	27511	13016	1.0621
AgeDriver:BodyCar	1	0.22	27510	13016	0.4264
VehValue:AgeDriver:BodyCar	1	0.01	27509	13016	0.0232
AgeCar:BodyCar	1	22.81	27508	12993	43.4419
VehValue:AgeCar:BodyCar	1	0.64	27507	12993	1.2161
AgeDriver:AgeCar:BodyCar	1	0.37	27506	12992	0.6985
VehValue:AgeDriver:AgeCar:BodyCar	1	0.95	27505	12991	1.8090
				Pr (>F)	
NULL					
VehValue		< 2.2e-16		***	
AgeDriver		< 2.2e-16		***	
VehValue:AgeDriver		1.785e-06		***	
AgeCar		0.0002758		***	
VehValue:AgeCar		< 2.2e-16		***	
AgeDriver:AgeCar		0.5239807			
VehValue:AgeDriver:AgeCar		0.6354142			
BodyCar		2.322e-05		***	
VehValue:BodyCar		0.3027417			
AgeDriver:BodyCar		0.5137532			

VehValue:AgeDriver:BodyCar	0.8790387
AgeCar:BodyCar	4.446e-11 ***
VehValue:AgeCar:BodyCar	0.2701387
AgeDriver:AgeCar:BodyCar	0.4032799
VehValue:AgeDriver:AgeCar:BodyCar	0.1786425

```
> ClaimsModgnm2<-update(ClaimsModgnm2, ~.-AgeDriver:AgeCar
+ -VehValue:AgeDriver:AgeCar-VehValue:BodyCar
+ -AgeDriver:BodyCar-VehValue:AgeDriver:BodyCar-VehValue:AgeCar:BodyCar
+ -AgeDriver:AgeCar:BodyCar-VehValue:AgeDriver:AgeCar:BodyCar)
> anova(ClaimsModgnm2, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F	Pr (>F)
NULL		27520	13514			
VehValue	1	366.42	27519	13147	697.950	< 2.2e-16 ***
AgeDriver	1	56.95	27518	13090	108.481	< 2.2e-16 ***
VehValue:AgeDriver	1	11.98	27517	13078	22.826	1.782e-06 ***
AgeCar	1	6.95	27516	13071	13.233	0.0002756 ***
VehValue:AgeCar	1	44.61	27515	13027	84.982	< 2.2e-16 ***
BodyCar	1	9.23	27514	13017	17.587	2.753e-05 ***
AgeCar:BodyCar	1	17.71	27513	13000	33.730	6.400e-09 ***

```
> GNMClaims2<-predict(ClaimsModgnm2, type="response")
> rc$GNMClaims2<-with(rc, predict(ClaimsModgnm2, type="response"))
```

```
#multivariate analysis with gaussian and GNM approach#
> rc$ACost<-with(rc, ClaimAmount/ClaimNb)
> CostModgnm2<-gnm(ACost~VehValue*AgeDriver*AgeCar*BodyCar,
+ weights=ClaimNb, data=rc, family=gaussian(link=log))
> anova(CostModgnm2, test="F")
```

	Df	Deviance	Resid.Df	Resid.Dev	F
NULL			27520	3.2697e+10	
VehValue	1	80689564	27519	3.2616e+10	68.3971
AgeDriver	1	32960850	27518	3.2583e+10	27.9395
VehValue:AgeDriver	1	36244011	27517	3.2547e+10	30.7225
AgeCar	1	24114	27516	3.2547e+10	0.0204
VehValue:AgeCar	1	2547704	27515	3.2545e+10	2.1596
AgeDriver:AgeCar	1	107169	27514	3.2545e+10	0.0908
VehValue:AgeDriver:AgeCar	1	3638873	27513	3.2541e+10	3.0845
BodyCar	1	17260580	27512	3.2524e+10	14.6311
VehValue:BodyCar	1	4991925	27511	3.2519e+10	4.2314
AgeDriver:BodyCar	1	2910036	27510	3.2516e+10	2.4667
VehValue:AgeDriver:BodyCar	1	150995	27509	3.2516e+10	0.1280
AgeCar:BodyCar	1	3577799	27508	3.2512e+10	3.0327
VehValue:AgeCar:BodyCar	1	2287881	27507	3.2510e+10	1.9393
AgeDriver:AgeCar:BodyCar	1	667589	27506	3.2509e+10	0.5659
VehValue:AgeDriver:AgeCar:BodyCar	1	681880	27505	3.2508e+10	0.5780
				Pr (>F)	
NULL					
VehValue		< 2.2e-16		***	
AgeDriver		1.261e-07		***	
VehValue:AgeDriver		3.004e-08		***	
AgeCar		0.886314			
VehValue:AgeCar		0.141695			
AgeDriver:AgeCar		0.763111			
VehValue:AgeDriver:AgeCar		0.079052		.	
BodyCar		0.000131		***	
VehValue:BodyCar		0.039691		*	
AgeDriver:BodyCar		0.116292			
VehValue:AgeDriver:BodyCar		0.720526			
AgeCar:BodyCar		0.081611		.	

```

VehValue:AgeCar:BodyCar          0.163752
AgeDriver:AgeCar:BodyCar         0.451904
VehValue:AgeDriver:AgeCar:BodyCar 0.447104

> CostModgnm2<-update(CostModgnm2, ~.-AgeCar-VehValue:AgeCar
+ -AgeDriver:AgeCar-AgeDriver:BodyCar-VehValue:AgeDriver:BodyCar
+ -VehValue:AgeCar:BodyCar-AgeDriver:AgeCar:BodyCar
+ -VehValue:AgeDriver:AgeCar:BodyCar-VehValue:AgeDriver:AgeCar
+ -BodyCar:AgeCar)
> anova(CostModgnm2, test="F")

Df Deviance Resid.Df  Resid.Dev      F      Pr(>F)
NULL                      27520 3.2697e+10
VehValue                  1 80689564  27519 3.2616e+10 68.2632 < 2.2e-16 ***
AgeDriver                  1 32960850  27518 3.2583e+10 27.8848 1.297e-07 ***
VehValue:AgeDriver        1 36244011  27517 3.2547e+10 30.6624 3.099e-08 ***
BodyCar                     1 18471166  27516 3.2529e+10 15.6266 7.736e-05 ***
VehValue:BodyCar           1  5254499  27515 3.2524e+10  4.4453  0.03501   *

```

```

> GNMCost2<-predict(CostModgnm2, type="response")
> rc$GNMCost2<-with(rc, predict(CostModgnm2, type="response"))

> #deterministic risk premium#
> rc$RiskPremium2<-with(rc, (GNMCost2*GNMClaims2)/Exposure)

> #stochastic risk premium with GNM approach#
> PRSModgnm2<-gnm(RiskPremium2~VehValue*AgeDriver*AgeCar*BodyCar,
+ weights=Exposure, data=rc, family=Gamma(link=log))
> GNMSRiskPremium2<-predict(PRSModgnm2, data=rc, type="response")
> rc$GNMSRiskPremium2<-with(rc, GNMSRiskPremium2)

```

The positive approach of multivariate analysis is that it investigates about correlation between variables, it is essentials for a complete analysis respect univariate approach that consider the assumption of independent variables.

```
> #Expected losses and actual losses
> expectedlosses<-with(rc, sum(GNMSRiskPremium2*Exposure) )
> expectedlosses
[1] 9361125
> actuallosses<-with(rc, sum(ClaimAmount) )
> actuallosses
[1] 9342125
> round((expectedlosses-actuallosses)/actuallosses,4) #difference#
[1] 0.002

> #fees#
> F<-20
> H<-0.25

#tariff requirement#
> GNMFabTariff2<-with(rc,sum(ClaimAmount) /
+ sum(GNMSRiskPremium2*Exposure))
> GNMFabTariff2
[1] 0.9979703

> #Commercial tariff#
> rc$GNMTariffPremium2<-with(rc,
+ (GNMSRiskPremium2*GNMFabTariff2+F) / (1-H) )

#Mean commercial tariff#
> MGNMTariffPremium2<-with(rc, mean(GNMTariffPremium2))
```

```
> MGNMTariffPremium2
[1] 234.6165

#Loss ratio#
> with(rc, sum(ClaimAmount) / sum(GNMTariffPremium2) )
[1] 1.446848
```

In the same way of GLM there are overestimations about expected losses that it reflects lower tariff requirement but the same mean collective tariff premium, not only there are lower number of risk coefficients but the intercept is rising up than univariate approach.

```
#Tariff coefficients#
> coefRiskPremium2<- (coef(PRSModgnm2))
> tableCoeff<-data.frame(coefficient=
+ c(levels<-names(coefRiskPremium2)),
+ coefficientsRiskPremium=exp(as.numeric(coefRiskPremium2)))
> print(tableCoeff,digits=2)
```

	coefficient	coefficientsRiskPremium
1	(Intercept)	265.06
2	VehValue	0.70
3	AgeDriver	1.00
4	VehValue:AgeDriver	1.06
5	AgeCar	1.00
6	VehValue:AgeCar	0.97
7	AgeDriver:AgeCar	1.00
8	VehValue:AgeDriver:AgeCar	1.00
9	BodyCar	0.94
10	VehValue:BodyCar	1.01
11	AgeDriver:BodyCar	1.00
12	VehValue:AgeDriver:BodyCar	1.00
13	AgeCar:BodyCar	1.01

```

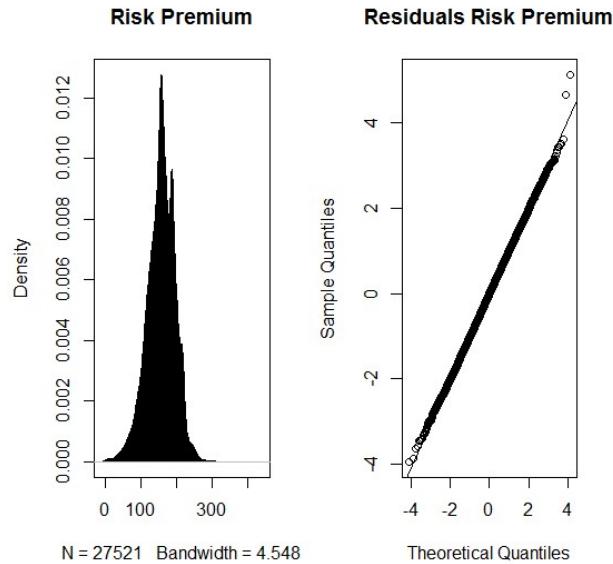
14          VehValue:AgeCar:BodyCar           1.00
15          AgeDriver:AgeCar:BodyCar         1.00
16          VehValue:AgeDriver:AgeCar:BodyCar 1.00

```

```

> #Distributions Risk Premium and residuals#
> par(mfrow=c(1,2))
> b<-density(GNMSRiskPremium2)
> plot(b,type="h", col="black",main="Risk Premium")
> x<-rnorm(resid(PRSModgnm2))
> qqnorm(x,main="Residuals Risk Premium")
> qqline(x)

```



Pic.7 - Distributions Risk Premium and residuals for multivariate GNM

```

#test on the model#
> summary(PRSModgnm2)

```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.649e-08	-9.820e-09	-6.386e-10	0.000e+00	0.000e+00

Coefficients:

	Estimate	Std.Error	t value	Pr(> t )
(Intercept)	5.579949	0.000000	5.823e+14	$<2e-16^{***}$
VehValue	-0.354753	0.000000	-6.808e+13	$<2e-16^{***}$
AgeDriver	-0.002626	0.000000	-9.880e+11	$<2e-16^{***}$
VehValue:AgeDriver	0.060037	0.000000	4.004e+13	$<2e-16^{***}$
AgeCar	0.003971	0.000000	1.092e+12	$<2e-16^{***}$
VehValue:AgeCar	-0.027218	0.000000	-1.658e+13	$<2e-16^{***}$
AgeDriver:AgeCar	0.000000	0.000000	9.280e-01	0.35361
VehValue:AgeDriver:AgeCar	0.000000	0.000000	2.865e+00	0.00418**
BodyCar	-0.059320	0.000000	-5.450e+13	$<2e-16^{***}$
VehValue:BodyCar	0.014059	0.000000	2.558e+13	$<2e-16^{***}$
AgeDriver:BodyCar	0.000000	0.000000	7.590e-01	0.44791
VehValue:AgeDriver:BodyCar	0.000000	0.000000	3.026e+00	0.00248**
AgeCar:BodyCar	0.007606	0.000000	1.801e+13	$<2e-16^{***}$
VehValue:AgeCar:BodyCar	0.000000	0.000000	2.898e+00	0.00376**
AgeDriver:AgeCar:BodyCar	0.000000	0.000000	-1.020e+00	0.30750
VehValue:AgeDriver:AgeCar:BodyCar	0.000000	0.000000	-2.418e+00	0.01563*

(Dispersion parameter for Gamma family taken to be 5.557383e-27)

Residual deviance: 3.6614e-14 on 27505 degrees of freedom

AIC: -1781693

Number of iterations: 1

```

> R2gnm2<-cor(rc$RiskPremium2,predict(PRSModgnm2, type="link"))^2
> R2gnm2
[1] 0.8704423
> BIC(PRSModgnm2)
[1] -1781553
> AIC(PRSModgnm2)
[1] -1781693
> deviance(PRSModgnm2)
[1] 3.661369e-14
> df.residual(PRSModgnm2)
[1] 27505

```

By last results there is a difference about -10% from deviance explained by multivariate to univariate approach. This result means that approach with assumptions of independent variables classified into homogeneous risk are better than multivariate approach without categorical variables. Also the first one seems enough because the portfolio is wide and moreover univariate win from comparing the deviance explained between these two type of approach without the use of categorized variables.

## Commercial MTPL Tariff with GAM Approach

Generalized Additive Models [S.N. Wood] extend Generalized Linear Models in the predictor  $\eta_i = \sum_{p=1}^n x_{ip}\beta_p + \sum_{j=1}^m f_j(x_{ij})$  made up by one parametric linear part and one non parametric part built by the sum of unknown “smoothing” functions of the covariates. For the estimators are used splines, functions made up by combination of little polynomial segment joined in knots with the target to have less deviance.

$$g(E[Y|X]) = \beta_0 + \sum_{j=1}^n x_{ij}\beta_j \Rightarrow g(E[Y|X]) = \beta_0 + \sum_{p=1}^n x_{ip}\beta_p + \sum_{j=1}^m f_j(x_{ij})$$

Will looking for a function with less deviance, so is added a smoothing parameter  $\lambda$  to control curve behavior, with target of little bending and little changing slope of the function used:

- With  $\lambda = 0$  function  $f$  hasn't constraints and it interpolates points;
- With  $\lambda = \infty$  function  $f$  is a straight line.

$$P_\lambda(f) = \sum_{i=1}^n \{Y_i - f(x_i)\}^2 + \lambda \int (f''(x))^2 dx$$

## Univariate Analysis

In univariate analysis are used tariff variables without cluster classification loaded in GLM and GNM models, because GAM model doesn't converge, but there are knots that represent where is located changing class, indeed number of knots are equal to the number of tariff classes inside each covariate.

For frequency model are used generic splines with quasi Poisson distribution.

```
> library(mgcv)
#univariate analysis with quasi Poisson and GAM approach#
> ClaimsModgam1<-gam(ClaimNb~s(AgeCar, k=4)+s(BodyCar, k=12)
+ +s(VehValue, k=30)+s(AgeDriver,k=6)
+ +offset(log(Exposure)),data=rc,family="quasipoisson"(link=log))
> summary(ClaimsModgam1)
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.577563	0.003653	-158.1	<2e-16 ***

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s (AgeCar)	2.910	2.993	16.64	4.35e-10 ***
s (BodyCar)	10.555	10.925	69.37	< 2e-16 ***
s (VehValue)	14.263	17.116	19.77	< 2e-16 ***
s (AgeDriver)	4.954	4.999	51.51	< 2e-16 ***

R-sq. (adj) = -2.29 Deviance explained = 7.77%

GCV = 0.45398 Scale est. = 0.43142 n = 27521

```
> GAMClaims1<-predict(ClaimsModgam1,type="response")
```

```
> rc$GAMClaims1<-with(rc,predict(ClaimsModgam1,type="response"))
```

Degrees of freedom for each splines are more than 1 and fewer than knots inside for each covariate, so is exact to put on splines for each variable and the model is unbiased.

For loss model and cost model are used cyclic cubic regression splines with Gamma distribution, the reason why is the start point is the same of the end point, this afford a control of the curve behavior and looking histogram of variables this is quite true for three covariates. In the cost model AgeCar variable is not removed because there aren't improvement.

```
#univariate analysis with Gamma and GAM approach#
> rc$ACost<-with(rc, ClaimAmount/ClaimNb)
> CostModgam1<-gam(ACost~s(AgeCar, bs="cc", k=4)
+ +s(BodyCar, bs="cc", k=12)+s(VehValue, bs="cc", k=30)
+ +s(AgeDriver, bs="cc", k=6), weights=ClaimNb,
+ data=rc,family="Gamma"(link=log))
> summary(CostModgam1)
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.58599	0.03486	160.2	<2e-16 ***

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(AgeCar)	1.668	2	1.109	0.260688
s(BodyCar)	7.247	10	2.117	0.002647 **
s(VehValue)	26.915	28	2.047	0.000435 ***
s(AgeDriver)	3.965	4	3.954	0.003077 **

R-sq. (adj) = 0.0065 Deviance explained = 1.48%

GCV = 9.2275 Scale est. = 39.307 n = 27521

```
> GAMCost1<-predict(CostModgam1,type="response")
> rc$GAMCost1<-with(rc,predict(CostModgam1,type="response"))
```

```
#deterministic risk premium#
> rc$RiskPremiumgaml<-with(rc, (GAMCost1*GAMClaims1)/Exposure)

#stochastic risk premium with GAM approach#
> PRSModgaml<-gam(RiskPremiumgaml~s(AgeCar, bs="cc", k=4)
+ +s(BodyCar, bs="cc", k=12)+s(VehValue, bs="cc", k=30)
+ +s(AgeDriver, bs="cc", k=6), weights=Exposure, data=rc,
+ family=Gamma(link=log))
> GAMSRiskPremium1<-predict(PRSModgaml,data=rc,type="response")
> rc$GAMSRiskPremium1<-with(rc, GAMSRiskPremium1)
```

Results are very good because expected losses are quite the same of actual losses and it afford a lower tariff requirement and lower mean commercial tariff premium despite of the other ones.

```
#Expected losses and actual losses
> expectedlosses<-with(rc, sum(GAMSRiskPremium1*Exposure))
> expectedlosses
[1] 9325999
> actuallosses<-with(rc, sum(ClaimAmount))
> actuallosses
[1] 9342125
> round((expectedlosses-actuallosses)/actuallosses,4) #difference#
[1] -0.0017

#fees#
> F<-20
> H<-0.25

#tariff requirement#
> GAMFabTariff1<-with(rc,sum(ClaimAmount) /
+ sum(GAMSRiskPremium1*Exposure))
```

```

> GAMFabTariff1
[1] 1.001729

#Commercial tariff#
> rc$GAMTariffPremium1<-with(rc,
+ (GAMSRiskPremium1*GAMFabTariff1+F) / (1-H) )

#Mean commercial tariff#
> MGAMTariffPremium1<-with(rc, mean(GAMTariffPremium1))
> MGAMTariffPremium1
[1] 232.8702

#Loss ratio#
> with(rc, sum(ClaimAmount) / sum(GAMTariffPremium1))
[1] 1.457698

#Tariff coefficients#
> coefRiskPremium1<-(coef(PRSModgam1))
> tableCoeff<-data.frame(coefficient=
+ c(levels<-names(coefRiskPremium1)),
+ coefficientsRiskPremium=exp(as.numeric(coefRiskPremium1)))
> print(tableCoeff,digits=2)

      coefficient   coefficientsRiskPremium
1          (Intercept)           149.526
2          s(AgeCar).1            0.995
3          s(AgeCar).2            1.079
4          s(BodyCar).1            0.833
5          s(BodyCar).2            1.175
6          s(BodyCar).3            1.130
7          s(BodyCar).4            1.147

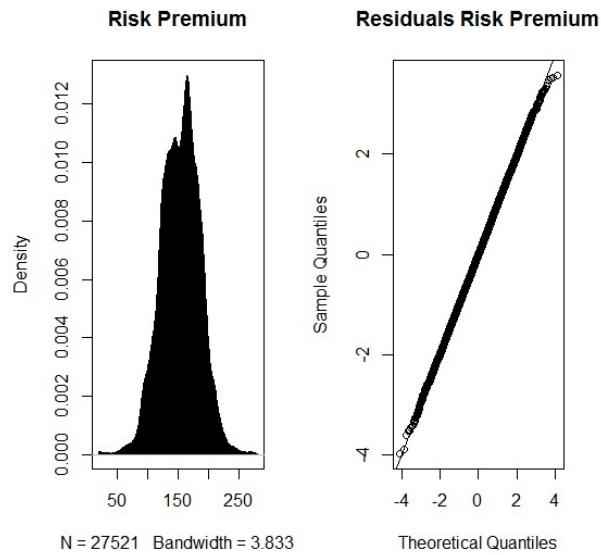
```

8	s (BodyCar) .5	0.820
9	s (BodyCar) .6	0.663
10	s (BodyCar) .7	1.010
11	s (BodyCar) .8	1.074
12	s (BodyCar) .9	1.147
13	s (BodyCar) .10	1.075
14	s (VehValue) .1	1.203
15	s (VehValue) .2	1.499
16	s (VehValue) .3	1.394
17	s (VehValue) .4	1.371
18	s (VehValue) .5	1.297
19	s (VehValue) .6	1.469
20	s (VehValue) .7	1.068
21	s (VehValue) .8	0.977
22	s (VehValue) .9	1.032
23	s (VehValue) .10	1.059
24	s (VehValue) .11	0.907
25	s (VehValue) .12	0.891
26	s (VehValue) .13	0.806
27	s (VehValue) .14	0.883
28	s (VehValue) .15	1.108
29	s (VehValue) .16	0.641
30	s (VehValue) .17	1.689
31	s (VehValue) .18	0.877
32	s (VehValue) .19	0.976
33	s (VehValue) .20	0.565
34	s (VehValue) .21	0.958
35	s (VehValue) .22	0.568
36	s (VehValue) .23	0.509
37	s (VehValue) .24	4.915
38	s (VehValue) .25	0.280

39	s (VehValue) .26	0.093
40	s (VehValue) .27	0.700
41	s (VehValue) .28	0.872
42	s (AgeDriver) .1	0.978
43	s (AgeDriver) .2	0.728
44	s (AgeDriver) .3	1.051
45	s (AgeDriver) .4	1.150

Other amazing results are the number of risk coefficients: 45 compared on 20 of the other ones. At the end the shape of the distribution has a regular behavior despite GLM and GNM.

```
#Distributions Risk Premium and residuals#
> par(mfrow=c(1,2))
> b<-density(GAMSRiskPremium1, from=20, to=280)
> plot(b,type="h", col="black",main="Risk Premium")
> x<-rnorm(resid(PRSModgam1))
> qqnorm(x,main="Residuals Risk Premium")
> qqline(x)
```



Pic.8 - Distributions Risk Premium and residuals for univariate GAM

```
> summary(PRSModgam1)
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.0074733	0.0002911	17201	<2e-16 ***

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(AgeCar)	1.999	2	9419	<2e-16 ***
s(BodyCar)	9.994	10	9463	<2e-16 ***
s(VehValue)	27.999	28	16914	<2e-16 ***
s(AgeDriver)	3.999	4	43712	<2e-16 ***

R-sq. (adj) = 0.957 Deviance explained = 96.2%

GCV = 0.00467 Scale est. = 0.0047664 n = 27521

#test on the model#

```
> pseudoR2gaml<-function(mod) {1-(deviance(mod)/mod$null.deviance)}
> pseudoR2gaml(PRSModgam1)
[1] 0.9618393
> AIC(PRSModgam1)
[1] 401048.2
> BIC(PRSModgam1)
[1] 401426.3
> deviance(PRSModgam1)
[1] 128.1023
> df.residual(PRSModgam1)
[1] 27476.01
```

The main characteristic of the GAM model is to have a lot of coefficients, so it is very good for tariff personalization.

## Multivariate Analysis

The analysis follows with multivariate approach using tensor smooths for each model because they are good for smooth interactions of quantities measured in different units and very different degrees of smoothness to different covariates.

```
#multivariate analysis with quasi Poisson and GAM approach#
> ClaimsModgam2<-gam(ClaimNb~te(AgeCar,BodyCar,VehValue,AgeDriver, k=4)
+ +offset(log(Exposure)),data=rc,family="quasipoisson"(link=log))
> summary(ClaimsModgam2)
```

Parametric coefficients:

	Estimate	Std.Error	t value	Pr(> t )
(Intercept)	-0.576709	0.003654	-157.8	<2e-16 ***

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
te(AgeCar,BodyCar,VehValue,AgeDriver)	103.2	120.6	20.68	<2e-16 ***
R-sq. (adj) = -2.26 Deviance explained = 8.3%				
GCV = 0.45372 Scale est. = 0.43143 n = 27521				

```
> GAMClaims2<-predict(ClaimsModgam2,type="response")
> rc$GAMClaims2<-with(rc,predict(ClaimsModgam2,type="response"))
```

For frequency model and loss model are put 4 knots because otherwise the model doesn't converge.

In the first one there are about 103 degrees of freedom, less than 256 = (4 variables^4 knots), so the model is correct.

```
#multivariate analysis with Gamma and GAM approach#
> rc$ACost<-with(rc, ClaimAmount/ClaimNb)
> CostModgam2<-gam(ACost~te(BodyCar, VehValue, AgeDriver, k=6)
+ +s(AgeCar, bs="cc", k=4), weights=ClaimNb,
```

```
+ data=rc, family="Gamma"(link=log))
> summary(CostModgam2)
```

Parametric coefficients:

	Estimate	Std.Error	t value	Pr(> t )
(Intercept)	5.47297	0.03178	172.2	<2e-16 ***

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
te(BodyCar, VehValue, AgeDriver)	168.060	176.2	2.608	<2e-16 ***
s(AgeCar)	1.577	2.0	0.812	0.355

R-sq. (adj) = -24.4 Deviance explained = 3.98%  
 GCV = 9.0793 Scale est. = 32.648 n = 27521

```
> CostModgam2<-gam(ACost~te(BodyCar, VehValue, AgeDriver, k=6),
+ weights=ClaimNb, data=rc, family="Gamma"(link=log))
> summary(CostModgam2)
```

Parametric coefficients:

	Estimate	Std.Error	t value	Pr(> t )
(Intercept)	5.47391	0.03178	172.2	<2e-16 ***

Approximate significance of smooth terms:

	Edf	Ref.df	F	p-value
te(BodyCar, VehValue, AgeDriver)	168.9	176.7	2.648	<2e-16 ***

R-sq. (adj) = -57.5 Deviance explained = 3.96%  
 GCV = 9.0802 Scale est. = 32.653 n = 27521

In the cost model are used 3 variables because by analysis of variance AgeCar is not relevant.

```
> GAMCost2<-predict(CostModgam2,type="response")
> rc$GAMCost2<-with(rc,predict(CostModgam2,type="response"))

> #deterministic risk premium#
> rc$RiskPremiumgam2<-with(rc, (GAMCost2*GAMClaims2)/Exposure)

#stochastic risk premium with GAM approach#
> PRSModgam2<-gam(RiskPremiumgam2~te(BodyCar,VehValue,AgeDriver,AgeCar,
+ k=4),weights=Exposure, data=rc, family="Gamma"(link=log))
> GAMSRRiskPremium2<-predict(PRSModgam2,data=rc,type="response")
> rc$GAMSRRiskPremium2<-with(rc, GAMSRRiskPremium2)
```

In the multivariate approach expected losses are overestimated so tariff requirement and loss ratio are lower than other approaches with rising mean commercial tariff premium.

```
> #Expected losses and actual losses
> expectedlosses<-with(rc, sum(GAMSRRiskPremium2*Exposure))
> expectedlosses
[1] 10867438
> actuallosses<-with(rc, sum(ClaimAmount))
> actuallosses
[1] 9342125
> round((expectedlosses-actuallosses)/actuallosses,4) #difference#
[1] 0.1633

#fees#
> F<-20
> H<-0.25
```

```
#tariff requirement#
> GAMFabTariff2<-with(rc,sum(ClaimAmount) /
+ sum(GAMSRiskPremium2*Exposure))
> GAMFabTariff2
[1] 0.8596438

#Commercial tariff#
> rc$GAMTariffPremium2<-with(rc,
+ (GAMSRiskPremium2*GAMFabTariff2+F) / (1-H) )

#Mean commercial tariff#
> MGAMTariffPremium2<-with(rc, mean(GAMTariffPremium2))
> MGAMTariffPremium2
[1] 248.5732

#Loss ratio#
> with(rc, sum(ClaimAmount) /sum(GAMTariffPremium2) )
[1] 1.365612

#Tariff coefficients#
> coefRiskPremium2<-coef(PRSModgam2)
> tableCoeff<-data.frame(coefficient=
+ c(levels<-names(coefRiskPremium2)),
+ coefficientsRiskPremium=exp(as.numeric(coefRiskPremium2)))
> print(tableCoeff,digits=2)

            coefficient   coefficientsRiskPremium
1                   (Intercept)      1.4e+02
2 te(BodyCar,VehValue,AgeDriver,AgeCar).1  1.2e-01
3 te(BodyCar,VehValue,AgeDriver,AgeCar).2  2.7e-02
4 te(BodyCar,VehValue,AgeDriver,AgeCar).3  3.7e+09
```

5	te(BodyCar, VehValue, AgeDriver, AgeCar)	.4	1.9e+00
6	te(BodyCar, VehValue, AgeDriver, AgeCar)	.5	6.6e+00
7	te(BodyCar, VehValue, AgeDriver, AgeCar)	.6	1.3e+04
8	te(BodyCar, VehValue, AgeDriver, AgeCar)	.7	4.6e-27
9	te(BodyCar, VehValue, AgeDriver, AgeCar)	.8	1.4e-01
10	te(BodyCar, VehValue, AgeDriver, AgeCar)	.9	6.5e+00
11	te(BodyCar, VehValue, AgeDriver, AgeCar)	.10	1.1e-12
12	te(BodyCar, VehValue, AgeDriver, AgeCar)	.11	2.0e-28
13	te(BodyCar, VehValue, AgeDriver, AgeCar)	.12	1.1e-01
14	te(BodyCar, VehValue, AgeDriver, AgeCar)	.13	1.8e+01
15	te(BodyCar, VehValue, AgeDriver, AgeCar)	.14	6.9e-04
16	te(BodyCar, VehValue, AgeDriver, AgeCar)	.15	3.7e-32
17	te(BodyCar, VehValue, AgeDriver, AgeCar)	.16	1.2e+00

---

127	te(BodyCar, VehValue, AgeDriver, AgeCar)	.126	3.2e-58
128	te(BodyCar, VehValue, AgeDriver, AgeCar)	.127	1.0e+67
129	te(BodyCar, VehValue, AgeDriver, AgeCar)	.128	2.1e+00
130	te(BodyCar, VehValue, AgeDriver, AgeCar)	.129	4.2e+00
131	te(BodyCar, VehValue, AgeDriver, AgeCar)	.130	5.2e-01
132	te(BodyCar, VehValue, AgeDriver, AgeCar)	.131	1.1e-01
133	te(BodyCar, VehValue, AgeDriver, AgeCar)	.132	2.8e+00
134	te(BodyCar, VehValue, AgeDriver, AgeCar)	.133	1.5e+00
135	te(BodyCar, VehValue, AgeDriver, AgeCar)	.134	3.9e+00
136	te(BodyCar, VehValue, AgeDriver, AgeCar)	.135	9.2e+00
137	te(BodyCar, VehValue, AgeDriver, AgeCar)	.136	1.3e+00
138	te(BodyCar, VehValue, AgeDriver, AgeCar)	.137	2.6e+00
139	te(BodyCar, VehValue, AgeDriver, AgeCar)	.138	1.0e+00
140	te(BodyCar, VehValue, AgeDriver, AgeCar)	.139	3.9e+00
141	te(BodyCar, VehValue, AgeDriver, AgeCar)	.140	4.8e-01
142	te(BodyCar, VehValue, AgeDriver, AgeCar)	.141	3.4e-01
143	te(BodyCar, VehValue, AgeDriver, AgeCar)	.142	8.3e-01

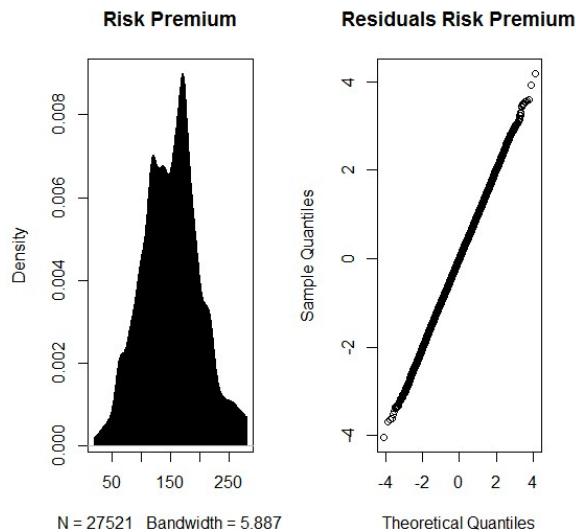
```

144 te(BodyCar,VehValue,AgeDriver,AgeCar).143      1.8e+00
-----
255 te(BodyCar,VehValue,AgeDriver,AgeCar).254      0.0e+00
256 te(BodyCar,VehValue,AgeDriver,AgeCar).255      1.9e-288

```

There are 256 risk coefficients and density distribution seems more enlarged.

```
#Distributions Risk Premium and residuals#
> par(mfrow=c(1,2))
> b<-density(GAMSRiskPremium2, from=20, to=280)
> plot(b,type="h", col="black",main="Risk Premium")
> x<-rnorm(resid(PRSModgam2))
> qqnorm(x,main="Residuals Risk Premium")
```



Pic.9 - Distributions Risk Premium and residuals for multivariate GAM

```
#test on the model#
> summary(PRSModgam2)
```

Parametric coefficients:

	Estimate	Std.Error	t value	Pr(> t )
(Intercept)	4.974757	0.002084	2387	<2e-16 ***

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
te(BodyCar, VehValue, AgeDriver, AgeCar)	254.9	255	184.7	<2e-16 ***

R-sq. (adj) = 0.693 Deviance explained = 84.8%  
 GCV = 0.2505 Scale est. = 0.24346 n = 27521

```

> pseudoR2gam2<-function(mod) {1-(deviance(mod)/mod$null.deviance)}
> pseudoR2gam2(PRSModgam2)
[1] 0.8482678
> AIC(PRSModgam2)
[1] 627073.6
> BIC(PRSModgam2)
[1] 629186.1
> deviance(PRSModgam2)
[1] 6766.419
> df.residual(PRSModgam2)
[1] 27265.09
  
```

From results of the stochastic risk premium degrees of freedom are lower than knots granted that the model is careful but is overestimated furthermore deviance explained is lower than univariate approach.

## Conclusions

		Mean commercial tariff	Tariff requirement	Loss Ratio	Residuals degrees of freedom	Expected Losses	Actual Losses	Explained Deviance	Risk coefficients
Uni- Variate Analysis	GLM	234,4587	1,000490	1,447822	27.501	9.337.547	9.342.125	96,96%	20
	GNM	234,4647	1,000476	1,447785	27.501	9.337.683	9.342.125	96,96%	20
	GAM	232,8702	1,001729	1,457698	27.476	9.325.999	9.342.125	96,20%	45
Multi- Variate Analysis	GLM	234,6486	0,9981246	1,446650	27.505	9.359.678	9.342.125	87,64%	16
	GNM	234,6165	0,9979703	1,446848	27.505	9.361.125	9.342.125	87,04%	16
	GAM	248,5732	0,8596438	1,365612	27.265	10.867.438	9.342.125	84,80%	256

At the end univariate approach looks like quite good to estimate these information. By these three models GAM results flexible to fit data with more risk coefficients so it's able to personalize commercial tariff but quite complex to realize, so it could be used in identifying the model before use GLM. This last one instead is user-friendly with faster elaboration but with poor flexibility to fit data, helps is input by cluster analysis that it granted more risk coefficients. GNM is a compromise to GLM because it affords some elaboration can't replicate with GLM and with lower values despite it; anova test support GNM.

## References

- [1]. E. Ohlsson, B. Johansson, (2010). "Non-Life Insurance pricing with Generalized Linear Models", Springer.
- [2]. G.A. Spedicato (2011). "Third party motor liability ratemaking with R", CAS Working Paper Disclaimer.
- [3]. S. N. Wood, (2006). "Generalized Additive Models – An Introduction with R", Chapman & Hall.
- [4]. C. Dutang, A. Charpentier, (2015). "CASdatasets Manual", <http://cas.uqam.ca/>
- [5]. H. Turner, D. Firth, (2007). "GNM a Package for Generalized NonLinear Models", R news, 7/2 8-12, R Foundation for Statistical Computing.
- [6]. Yu-Wei, Chiu (David Chiu), (2015). "Machine Learning with R Cookbook", Packt Publishing.