ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume 62 41 Number 2, 2014
http://dx.doi.org/10.11118/actaun201462020383

GENERALIZED LINEAR MODELS IN VEHICLE INSURANCE

Silvie Kafková[1], Lenka Křivánková[2]

[1] Masaryk University, Faculty of Economics and Administration, Lipová 41a, 602 00 Brno, Czech Republic
[2] Masaryk University, Faculty of Science, Kotlářská 2, 611 37 Brno, Czech Republic

Abstract

KAFKOVÁ SILVIE, KŘIVÁNKOVÁ LENKA. 2014. Generalized Linear Models in Vehicle Insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis,* 62(2): 383–388.

Actuaries in insurance companies try to find the best model for an estimation of insurance premium. It depends on many risk factors, e.g. the car characteristics and the profile of the driver. In this paper, an analysis of the portfolio of vehicle insurance data using a generalized linear model (GLM) is performed. The main advantage of the approach presented in this article is that the GLMs are not limited by inflexible preconditions. Our aim is to predict the relation of annual claim frequency on given risk factors. Based on a large real-world sample of data from 57 410 vehicles, the present study proposed a classification analysis approach that addresses the selection of predictor variables. The models with different predictor variables are compared by analysis of deviance and Akaike information criterion (AIC). Based on this comparison, the model for the best estimate of annual claim frequency is chosen. All statistical calculations are computed in R environment, which contains stats package with the function for the estimation of parameters of GLM and the function for analysis of deviation.

Keywords: vehicle insurance, generalized linear model, poisson distribution, link function, analysis of deviance, Akaike information criterion

1 INTRODUCTION

Actuarial science is a dynamically developing field dealing with an assessment of risk in insurance. Vehicle insurance is an insurance designed for cars, trucks, motorcycles, and other road vehicles. It is used to provide financial protection against the damage of the vehicle and a bodily injury resulting from traffic collisions. Moreover, it hedges against the liability which could arise in a traffic accident. The specific terms for vehicle insurance and its type vary with legal regulations. A review of actuarial modeling in vehicle insurance is given in Denuit (2007).

The process, by which insurers determine whether to insure an applicant and which premium to charge, is called vehicle insurance risk selection. The insurance premium is usually derived from an annual frequency of claims, which is modeled by using statistical data. This approach for computation of the premium can be found in Kaas (2009) and Ohlsson and Johansson (2010). The annual frequency of the claims is calculated from the number of the claims on a contract. They depend on many factors that are believed to have an impact on the expected cost of future claims. Those factors can include the car characteristics (vehicle body, vehicle age) and the profile of the driver (age, gender, driving history). Based on the idea of Heller and Jong (2008) and Kaas (2009), we develop models for the vehicle insurance.

The number of claims is a random variable. Based on Pearson's chi-squared test we assume that the number of claims on a contract is Poisson distributed.

Policyholders are divided into several tariff groups. For each group, generally different expected values of claims are assumed. The expected value of a Poisson distributed random variable is equal to its variance. Because such expected values in individual groups are different, it leads to the occurrence of heteroskedasticity. Therefore, we cannot usethe classical linear regression model.

The expected value of the random variable with the Poisson distribution is always positive. Under the assumption that the mean depends linearly

on the explanatory variables, its positivity cannot be guaranteed. Therefore, logarithmic transformation guaranteeing the positivity is used. In such case, instead of obtaining an additive model we get a model with a multiplicative effect on the mean.

These are the reasons that lead us to use generalized linear models (GLMs). In particular, we use GLM with the Poisson distributed response variable and with the logarithmic link function. Generalized linear models became very popular since their introduction in Nelder and Wedderburn (1972), primarily due to the ability to handle discrete data via an extension of the familiar Gaussian regression model to the models based on underlying exponential family of distributions. A basic notation, definition and framework of GLMs are described e.g. in Dobson (2002). For the wide overview on GLMs see the standard text McCullagh and Nelder (1989).

The aim of this paper is to develop a suitable model for an annual frequency of claims. Based on this model, the actuary can determine an adequate insurance premium for each group of drivers. The analysis of deviance and the Akaike information criterion are used for comparison of the examined models.

Over the last years generalized linear models became a favorite statistical tool to model actuarial data. We refer to Haberman and Renshaw (1996) for an overview of applications of GLMs in actuarial science. For example, Gschlössl, Schoenmaekers and Denuit (2011) described the application of GLM for a construction of life tables in life insurance. Another application of GLMs in life insurance is introduced in Cerchiara, Edwards and Gambini (2008), where GLMs are used in context of lapse risk as a mean to understand the relationship between risk factors and to calibrate the lapse risk as accurately as possible. Advantages of the GLMs approach are discussed in Antonio and Beirlant (2007). Furthermore, they presented the usage of generalized linear mixed models in actuarial mathematics.

The insurance portfolios have very specific characteristics, because, for many policies, there are no claims observed in the insurance history for a given period. It means that the data contains lots of zeros and therefore the GLMs may not give satisfactory results. This common situation considering the insurance data is discussed in Wolny-Dominiak (2012).

2 MATERIALS AND METHODS

The most common approach for modeling the relationships between variables uses linear regression models. A disadvantage of the standard linear regression model is the assumption of normally distributed observations, which does not allow appropriate modeling of counts, frequencies, binary or skewed data.

Another assumption of the linear regression models is that the mean of observations is a linear function of the parameters. Accordingly, it permits only additive models but not multiplicative ones. Moreover, linear regression models assume an independence of the variance and the mean while the variability often increases with the mean value in real data.

If the data does not comply to the above mentioned properties of linear regression model, we can use generalized linear models which do not require such strict assumptions.

2.1 Generalized Linear Models

In this section, we give only summary of the main characteristics of generalized linear models (GLMs). For a broad introduction to the generalized linear models, we refer to McCullagh and Nelder (1989), Dobson (2002) and Hardin and Hilbe (2007). Main attributes of the GLMs are the generalization of probability distribution of the dependent variable and giving a possibility to transform the data.

GLMs extend the framework of linear regression models with normal distribution to the class of distributions from the exponential family. It allows modeling of large numbers of types of variables (counts, frequencies, etc.) and to treat skewed probability distributions of the data, too. Densities from the exponential family are defined in the following canonical form.

Definition 1. The set of probability density functions (p.d.f.) written of the form

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

is called the *exponential family*, where θ and ϕ are parameters, θ is called the canonical parameter or the scale parameter and ϕ the dispersion parameter. a(θ) and $c(y, \phi)$ are known functions determining the actual probability function such as Binomial, Poisson, Normal or Gamma.

Further generalization uses a link function which allows to model transformed data. The link function makes a connection between the mean and a linear function of the explanatory variables. A transformation of the mean is modeled as a linear function of explanatory variables.

Definition 2. The *link function* $g(\mu)$ is a monotonic differentiable function of the form

$$g(\mu)=\mathbf{x}'\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the vector of regression parameters and $\mathbf{x}$ is a vector of the explanatory variables.

The link function $g(\mu)$ determines how the mean is related to the explanatory variables $\mathbf{x}$. Common link functions $g(\mu)$ are given in the Tab. I, in relation with specific probability distributions of the data.

I: *Commonly used link functions*

Distribution	Link function	$g(\mu)$
normal	identity	μ
poisson	log	$\ln(\mu)$
binomial	logit	$\ln\left(\frac{\mu}{1-\mu}\right)$
	cloglog	$\ln\left(-\ln\left(1-\frac{\mu}{n}\right)\right)$
exponential	log	$\ln(\mu)$

Definition 3. Let Y be a random variable with mean denoted by μ and p.d.f. from the exponential family. Then the *generalized linear model (GLM)* is given by

$$g(\mu) = \mathbf{x}'\boldsymbol{\beta},$$

where $g(\mu)$ is the link function.

The generalized linear models provide relatively simple and robust way to analyze the effect of many different factors on some observed event. The GLMs are used for valuation of insurance policies due to the number of claims. In GLM, it is assumed that the number of the claims is a dependent variable which follows Poisson distribution and which depends on known predictors. The predictors characterize the insured individual or vehicle, e.g. gender, age, engine capacity.

2.2 Methods of Comparing Different Models

Determining appropriate model is the basis of regression modeling. One important principle of regression modeling is the principle of simplicity. The simpler model, well describing the data, gets priority over the more complex model that describes the data almost perfectly.

2.2.1 Analysis of Deviance

Along with the basic generalized linear model, we also take into account the following partial models, which are called submodels.

Definition 4. The *full model*, denoted as GLM_{max}, satisfies the following conditions:
- it has the same distribution as the proposed model,
- it has the same link function as the proposed model,
- the number of parameters is equal to the number of the response variables.

The response variables are determined by the full model with residues equal to zero.

Definition 5. The *null model*, denoted as GLM_{min}, satisfies the following condition:
- it has the same distribution as the proposed model,
- it has the same link function as the proposed model,
- the number of parameters is equal to one.

The full model is an indicator of the "best regression" and the null model gives the "worst regression" with given distribution and link function. The proposed model will be somewhere between these two extreme models. The relevance of the proposed model will be evaluated via comparison with these models. However, such comparison is permitted only for the original model and its submodel.

Definition 6. Consider GLM with design matrix $\mathbf{X}_{\mathbf{n\times m}}$ and vector of parameters $\boldsymbol{\beta}_{\mathbf{m}}$. Its *submodel*, denoted as GLM_{sub}, with design matrix $\mathbf{Q}_{\mathbf{n\times q}}$ and vector of parameters $\boldsymbol{\beta}_{\mathbf{q}}$ satisfies the following conditions:
- it has the same distribution as the proposed GLM,
- it has the same link function as the proposed GLM,
- the number of parameters is $q < m$ and columns of the design matrix $\mathbf{Q}_{\mathbf{n\times q}}$ are linear combinations of columns of the design matrix $\mathbf{X}_{\mathbf{n\times m}}$.

The deviance is defined as a measure of distance between the full model and the proposed submodel.

Definition 7. The *deviance*, denoted as *dev*, is given by

$$dev = 2(l_{max} - l),$$

where l_{max} is logarithm of the likelihood function of the full model and l is logarithm of the likelihood function of the proposed submodel.

The deviances of the models are used for their comparison as described in the next theorem.

Theorem 8. *Consider GLM with vector of parameters* $\boldsymbol{\beta}_{\mathbf{m}}$ *and its submodel* GLM_{sub} *with* $\boldsymbol{\beta}_{\mathbf{q}}$, *where* $q < m < n$. *If the submodel* GLM_{sub} *is suitable, then the difference of deviances*

$$\Delta dev = dev_{sub} - dev$$

fulfills asymptotically χ^2 *distribution with* $(m - q)$ *degrees of freedom.*

For more details see Kaas (2009). Hence, for $\Delta dev > \chi^2_{1-\alpha}(m - q)$ we reject the assumption that the submodel is suitable.

2.2.2 Information Criteria

There are no perfect models. The idea is to find a model which is the best approximation of reality. We try to minimize the loss of information. The information criteria indicate that information is lost, when a model is used to describe the reality. They balance between accuracy of fitting the data and complexity of the model. Information criteria for selecting the minimal „good" model are for example:
- Akaike information criterion (AIC),
- Schwarz-Bayesian information criterion (BIC),
- Hannan-Quinn information criterion,
- Deviance information criterion (DIC).

In our case study, Akaike information criterion will be used for comparison of the models.

Definition 9. *Akaike information criterion (AIC)* is given as

$$AIC = -2l + 2k,$$

where k is the number of model parameters and l is logarithm of the likelihood function of the proposed model.

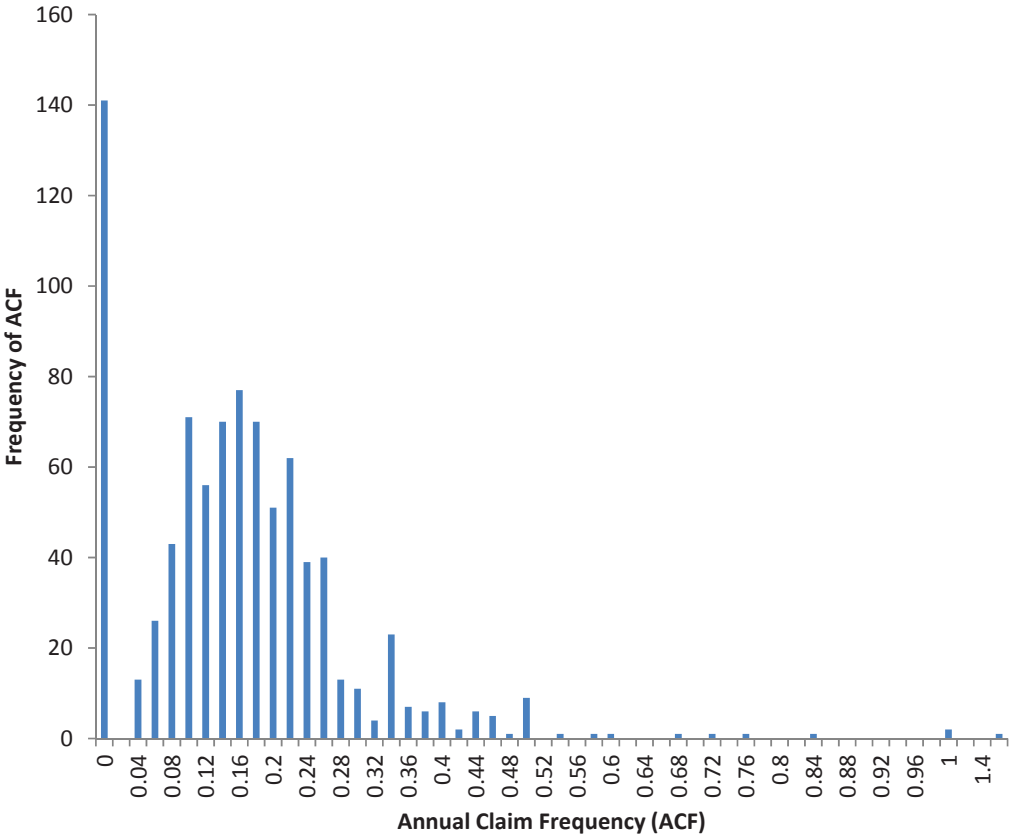
Preferred model is considered to be that with the lowest AIC.

3 RESULTS AND DISCUSION

Every person, when applying for vehicle insurance policy, is assigned to a class, that is homogeneous in terms of risk. One of the criteria used for assigning an individual to a certain class is the number of claims. Thus, it is very important task for insurance companies to model the number of claims in a given insurance portfolio.

Our aim is to predict relation of annual claim frequency on given risk factors. A data set from vehicle insurance will be processed. The data for our case study can be found in (Heller and Jong, 2008). The data set is based on one-year vehicle insurance policies recorded in 2004 or 2005. There are 57 410 policies and 3 913 of them (6.82%) have at least one claim. The total amount of claims is 4 176. We see, that the histogram of annual claim frequency is strongly right-skew (Fig. 1).

The GLMs are suitable for analysis of non-normal data, i.e. insurance data. Necessary procedures



1: *Histogram of Annual Claim Frequency*

II: *Variables in a data set*

Notation	Name of Variable	Range
expo	Exposure	0–1
clm	Claim occurrence	0 (no), 1 (yes)
numclaims	Number of claims	0, 1, 2, …
veh_body	Vehicle body type	hatchback, sedan, station wagon
veh_age	Vehicle age	1 (new), 2, 3, 4
area	Area of residence	A, B, C, D, E, F
gender	Gender	male, female
agecat	Age band of policyholder	1 (youngest), 2, 3, 4, 5, 6

III: *The estimate of parameters*

Coefficients					
(Intercept)	veh_body2	veh_body3	veh_age2	veh_age3	veh_age4
–1.62140	0.06056	0.09926	0.05294	–0.07510	–0.12863
area2	area3	area4	area5	area6	gender2
0.04384	0.01784	–0.11371	–0.01835	0.11873	–0.01132
agecat2	agecat3	agecat4	agecat5	agecat6	
–0.16865	–0.23317	–0.23977	–0.45026	–0.45845	

are implemented in R software environment. All variables in data set are given in Tab. II.

The drivers can be divided into groups on the basis of the risk factors (gender, age category, area, vehicle body, vehicle age). From these five risk factors and their values we get 864 groups. For each group, the total amount of exposure during the year (expo) and total of claims (numclaims) are known. We model the average number of claims per contract (numclaims/expo).

We try to find well-fitting GLM for the claim frequency in terms of the risk factors. For the number of claims per contract, it is reasonable to assume Poisson distribution. Our first GLM for data fitting is a model from Poisson family with log-link, which parameters are predicted in Tab. III. The coefficients are given relatively with respect to the standard class (veh_body1, veh_age1, area1, gender1, agecat1). The coefficients are taken to be zero for the standard class.

According to predicted parameters, the best group is the one with veh_body1, veh_age4, area4, gender2 and agecat6. The corresponding average number of claims equals to

$$e^{(-1.62140-0.12863-0.11371-0.01132-0.45845)} = 0.097.$$

That means one claim per 10.3 years on average.

3.1 Comparison of the Models

In this subsection, the models with different risk factors are compared. In the following Tab. IV we test the null hypothesis that adding a risk factor to our preceding model actually has no effect. The deviance (dev) for assessing the suitability of the proposed submodel is used. We assume the difference in deviance (Δdev) between the preceding model and the proposed model has χ^2 distribution with Δdf degrees of freedom. This is given in Theorem 8, where preceding model is a submodel of the proposed model.

According to the analysis of deviance, the best model is 1+agecat+veh_age. However, we choose the model 1+agecat+veh_age+area, although

$$\Delta dev = 10.58 < \chi^2_{0.95}(5) = 11.07.$$

The test of the statistic is close to the critical value of χ^2 distribution. We can support the inclusion of the parameter area by calculation of AIC.

Although the AIC penalizes the number of parameters, the selected model has smaller AIC than its submodel, for the model 1+agecat+veh_age it is AIC = 127 900 and for the model 1+agecat+veh_age+area it holds AIC = 127 200. Hence, according to AIC, the model is improved. Furthermore, based on an educated guess, significance of the factor area is not negligible.

4 CONCLUSION

Considering that real data from vehicle insurance is not normally distributed, we cannot use the standard linear regression model. This paper proposes an estimate of annual claim frequency in vehicle insurance based on General Linear Model. It represents a work devoted to better understanding, using data of vehicle insurance, and how GLMs can be used to explain the relation of annual claim frequency on given risk factors.

The case study results confirm the importance of three factors: age group of policy holder (agecat), vehicle age (veh_age) and area of residence (area). This particular case study shows that the gender

IV: *The table of analysis of deviance*

Model specification	df	dev	Δdev	Δdf
1	855	1048.0		
1+*veh_body*	853	1043.3	4.71	2
1+*veh_age*	852	1027.0	20.99	3
1+*area*	850	1033.2	14.82	5
1+*gender*	854	1047.7	0.38	1
1+*agecat*	850	972.7	75.33	5
1+agecat+*veh_age*	847	953.5	19.16	3
1+agecat+veh_age+*area*	842	942.9	10.58	5

or vehicle body type (veh_body) are relatively unimportant for annual claim frequency.

When the model was being created, we also had in mind the principle of simplicity. Therefore, we used analysis of deviance to compare relevance of the submodels. Our proposed model is quite simple, which is important for its use in practice.

SUMMARY

The aim of this paper is to estimate an annual frequency of claims (AFC) from which the premium in vehicle insurance is derived. It is considered that the AFC depends on many risk factors. We take into account these five factors – vehicle body type, vehicle age, area of residence, gender of policyholder and age band of policyholder. The generalized linear models (GLMs) are used for the estimation of AFC in this paper. This approach is compared with commonly used linear regression and the advantages of GLMs are shown. In the initial section, the framework of GLMs is introduced, including the definition of the link function and the specification of the exponential family of probability density functions. Then methods of model comparison, analysis of deviance and minimization of the information loss, are shown. The main part of the paper consists of a case study, where the GLMs are applied in vehicle insurance. We process a data set based on 57 410 one-year vehicle insurance policies. The drivers are divided into groups on the basis of the risk factors. For each group, we model the average number of claims per contract. The aim is to find a well-fitting GLM for the claim frequency in terms of the risk factors. The Poisson distribution is assumed for the number of claims per contract and the log-link function is used. Several different models containing various risk factors are considered. Analysis of deviance, based on a comparison of the goodness of fit, is used to select the best model. According to the analysis of deviance, the suitable model has two risk factors (age group of policyholder and vehicle age). Nevertheless, the more complex model with one other factor (area of residence) is taken into account. The significance of the factor area is not negligible in practice. Although the significance of the risk factor area is rejected at a significance level of 0.05, it is not rejected at a significance level of 0.1. The choice of the model with the factor area is substantiated by the calculation of Akaike information criterion (AIC), which is based on minimizing the information loss. The approach introduced in this paper proves to be a useful way to implement methodics of the generalized linear models into actuarial science.

REFERENCES

ANTONIO, K., BEIRLANT, J., 2007: Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics,* 40, 1: 58–76. ISSN 0167-6687.

CERCHIARA, R., EDWARDS, M., GAMBINI A., 2008: Generalized linear models in life insurance: decrements and risk factor analysis under Solvency II. In: *18th International AFIR Colloquium.* Available online: http://www.actuaries.org/index.cfm?lang=EN&DSP=AFIR&ACT=COLLOQUIA.

DENUIT, M. et al., 2007: *Actuarial modelling of claim counts: risk classification, credibility and bonus-malus systems.* Hoboken: Wiley, 384 p. ISBN 978-0-470-02677-9.

DOBSON, A. J., 2002: *An Introduction to Generalized Linear Models.* Boca Raton: CRC Press, 225 p. ISBN 1-58488-165-8.

GSCHLÖSSL, S., SCHOENMAEKERS, P., DENUIT, M., 2011: Risk classification in life insurance: methodology and case study. *European Actuarial Journal,* 1, 1: 23–41. ISSN 2190-9733.

HABERMAN, S., RENSHAW, A. E., 1996: Generalized linear models and actuarial science. *The Statistician,* 45, 4: 407–436. ISSN 0039-0526.

HARDIN, J. W., HILBE, J., 2007: *Generalized linear models and extensions.* Texas: Stata Press, 387 p. ISBN 1597180149.

HELLER, G. Z., JONG P., 2008: *Generalized Linear Models for Insurance Data*. New York: Cambridge University Press, 204 p. ISBN 0521879140.

KAAS, R. 2009: *Modern Actuarial Risk Theory: Using R*. Heidelberg: Springer, 400 p. ISBN 9783642034077.

MCCULLAGH, P., NELDER, J. A., 1989: *Generalized Linear Models.* London: Chapman and Hall, 532 p. ISBN 0412317605.

NELDER, J. A., WEDDERBURN, R. W. M., 1972: Generalized linear models. *Journal of the Royal Statistical Society: Series A (General),* 135, 3: 370–384. ISSN 0035-9238.

OHLSSON, E. JOHANSSON, B., 2010: *Non-life insurance pricing with generalized linear models.* Berlin / Heidelberg: Springer, 174 p. ISBN 978-3-642-10790-0.

WOLNY-DOMINIAK, A., 2012: Modeling of claim counts using data mining procedures in R CRAN. In: RAMÍK, J. and STAVÁREK, D. (eds.) *Proceedings of 30th International Conference Mathematical Methods in Economics.* Karviná: Silesian University, School of Business Administration, 980–985. ISBN 978-80-7248-779-0.

Contact information

Silvie Kafková: 175424@mail.muni.cz
Lenka Křivánková: 142474@mail.muni.cz