

MODELLING CLAIM FREQUENCY IN VEHICLE INSURANCE

Jiří Valecký¹

¹ Department of Finance, Faculty of Economics, VŠB-TU Ostrava, Sokolská tř. 33, 701 21 Ostrava, Czech Republic

Abstract

VALECKÝ JIŘÍ. 2016. Modelling Claim Frequency in Vehicle Insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 64(2): 683–689.

The paper is focused on modelling claim frequency and extends the work of Kafková and Křivánková, 2014 (Kafková, S., Křivánková, L. 2014. Generalized linear models in vehicle insurance. *Acta universitatis agriculturae et silviculturae mendelianae brunensis*, 62(2): 383–388). We showed that overdispersion, non-linear systematic component and interacted rating factors should be considered when the claim frequency is modelled. We detected overdispersion in the Poisson model and employed the negative-binomial model to show that considering heterogeneity over insurance policies yields better fit of the model. We also analysed the linear effect of continuous rating factors and their mutual influences. We showed that non-linearity and interactions between rating factors yield the better fit of the model, as well as new findings related to the analysis of claim frequency. All empirical models were estimated on the insurance portfolio of Czech insurance company collected during the years 2004–2008.

Keywords: claim frequency, generalized linear models, heterogeneity, negative-binomial regression, overdispersion, Poisson regression, vehicle insurance

INTRODUCTION

In the vehicle insurance, differentiating premiums have been used in recent decades. The insurers usually set the annual premium to comply with the volume of the engine and the size of the district where the policyholder lives, although some insurers also consider the client's age. Thus, the annual premium is determined increasingly according to the risk undertaken or the policyholder's risk behaviour (risk profile) which yields an insurance claims.

Due to this current trend in insurance, generalized linear models (GLMs) have become a popular statistical tool to analyse and model claim frequency and severity. McCullagh and Nelder (1989) introduced the concept of GLMs and many others have developed this concept further. The first regression analysis using individual rating factors and also one of the first separate analyses of claim frequency–severity appeared in Kahane and Levy (1975), while the first application of GLMs was used to model the claim frequency for marine insurance in McCullagh and Nelder (1983) and the claim size

for motor insurance in McCullagh and Nelder (1989).

More applications of GLMs occurred mostly after the 1990s, when the insurance market was being deregulated in many countries and the GLMs were used to undertake a tariff analysis, for example in Brockman and Wright (1992), Andrade-Silva (1989) or Renshaw (1994), or to set premiums, for example in Murphy, Brockman and Lee (2000), Zaks, Frostig and Levikson (2006) or Branda (2014).

The linear regression model is not possible to apply and the Poisson regression model is used as a fundamental model for claim counts. Thus, to identify the relevant rating factors to obtain a well fitted model is crucial as shown in Kafková and Křivánková (2014). However, to obtain a good model, it is necessary to take into account some modelling issues.

Firstly, converting continuous variables into categorical factors simplifies the analysis and interpretation of the results. Even if quantiles are used to determine the cutpoints, such model incurs information loss available in the data and, in addition, is unrealistic because policyholders

close to but on opposite sides of the cutpoint are treated as very different rather than very similar. Disadvantages of grouping were described by Altman *et al.* (1994) and many others.

Secondly, the observed claim frequency experiences overdispersion and some type of mixed Poisson model is needed to apply. Next, the relationship between the outcome and individual characteristics is expressed via a link function, which transforms the outcome into a linear combination of given individual rating factors. This can be observed that this combination is nonlinear which may obtain incorrect assessments of the impact of rating factor on the claim frequency. Generalized additive models (GAMs) are a good choice for modelling nonlinearity (see Wood (2006) for a good introduction or Ohlsson and Johansson (2010), chap. 5) for details of the application of GAMs to claim frequency models in non-life insurance). Finally, another modelling issue is related to the presence of “modifiers,” i.e., variables that modify the relationships between the outcome and other variable via specific interaction.

Thus, we extended the work of Kafková and Křivánková (2014) and we showed that the overdispersion, non-linear systematic component and interacted rating factors should be considered when the claim frequency is modelled. The remainder of this paper is organized as follows. The methodology used in this study, as well as the issue of modelling claim frequency is described in Section 2. The results are presented in Section 3 and Section 4 gives the conclusions of this paper.

Modelling Claim Frequency

The concept of GLMs in actuarial theory is already well introduced, e.g. Jong and Heller (2008), Ohlsson and Johansson (2010). Therefore, we present only the basic framework in this section.

Exponential Dispersion Model

The key concept of all GLM models is the exponential dispersion model, which uses a probability density function (pdf) with the form of

$$f(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right],$$

where y is the response, θ is the canonical (natural) parameter or link function, $b(\theta)$ is the cumulant, ϕ is the dispersion parameter, and $c(y, \phi)$ is a normalization term, which guarantees that the probability function sums to unity. The choices of $b(\theta)$ and θ determine the actual probability function and GLM model. $b(\theta)$ is assumed to be twice continuously differentiable with an invertible first derivative. The first and second derivative with respect to θ gives the mean response μ and its variance, and thus

$$E[y] = b'(\theta) = \frac{\partial b(\theta)}{\partial \theta} = \mu,$$

$$V(y) = \phi b''(\theta) = \phi \frac{\partial b'(\theta)}{\partial \theta} = \phi \frac{\partial \mu}{\partial \theta} = \phi V(\mu),$$

where V is the variance function. However, the dispersion parameter ϕ is the same for all of the policyholders, while the parameter θ is allowed to depend on individual rating factors and it can be expressed as the systematic components η_i using the expression

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_j x_{ij} = \mathbf{x}_i \mathbf{b} = g(\mu_i),$$

where x_{ij} is the j th rating factor of the i th policyholder, β are unknown parameters to be estimated, $g(\cdot)$ is a known link function, and the inverse link function is the mean function $\mu_i = g^{-1}(\eta_i)$.

To obtain the parameter estimates, we may employ two estimators: iteratively re-weighted least squares based on Fisher scoring or the maximum likelihood Newton-Raphson type algorithm. The dispersion parameter ϕ can be estimated using the maximum likelihood method by solving $\partial l / \partial \phi = 0$, or by methods of moments on the basis of Pearson $\phi_x = X^2(N - r)^{-1}$ or deviance residuals $\phi_D = D(N - r)^{-1}$, where r is the number of parameters (including constants).

Fractional Polynomials

The expression $\mathbf{x}_i \mathbf{b} = g(\mu_i)$ is not necessarily linear. One of the techniques used to handle nonlinearity involves fractional polynomials (FPs). Thus, let us define the nonlinear function and rewrite the expression $g(\mu_i)$ in the form of

$$g(\mu_i) = \beta_0 + \sum_{j=1}^J \beta_{1j} F_j(x_{i1}) + \beta_2 x_{i2} + \dots + \beta_K x_{iK},$$

where $F_j = (x_{i1})^{p_j}$ is a particular type of power function. The power p_j could be any number, but Royston and Altman (1994) restricts the power in the set $S \in \{-2; -1; 0.5; 0; 0.5; 1; 2; 3\}$, where 0 denotes the log of the variable. The remaining functions are defined as

$$F_j(x_{i1}) = \begin{cases} x_{i1}^{p_j}, & p_j \neq p_{j-1}, \\ F_{j-1}(x_{i1}) \ln(x_{i1}), & p_j = p_{j-1}. \end{cases}$$

Finally, we note that the identification and comparison of the most appropriate FPs could be performed using the sequential procedure (Royston and Altman, 1994) or the closed test procedure (Marcus, Peritz and Gabriel, 1976), where the latter is generally preferred.

Mixed Poisson Model

The Poisson model applied to claim frequency often experiences overdispersion, which is explained by the heterogeneity of the mean over the population of policyholders and can be examined using score test, LM test or boundary LR test. Therefore we compared the Poisson model with

negative-binomial model derived as a Poisson-gamma mixed model with pdf

$$f_y(y) = \frac{\Gamma(1/\kappa + y_i)}{\Gamma(1/\kappa)\Gamma(y_i + 1)} \left(\frac{1}{\kappa\mu_i + 1} \right)^{1/\kappa} \left(\frac{\kappa\mu_i}{\kappa\mu_i + 1} \right)^{y_i},$$

where κ is negative binomial heterogeneity or overdispersion parameter, y_i is the observed claim frequency and μ_i is the mean response.

The canonical link function of this model is in the form of

$$g(\mu_i) = \ln \frac{\kappa\mu_i}{\kappa\mu_i + 1} = \mathbf{x}_i\boldsymbol{\beta}.$$

However, it is used rarely in practice because of the need to know κ , while the condition $\mu_i > 0$ implies that $\mathbf{x}_i\boldsymbol{\beta} < 0$ and beta restrictions are necessary. Therefore, we use a non-canonical log link function, $g(\mu_i) = \ln(\mu_i) = \mathbf{x}_i\mathbf{b}$ with the inverse function $\mu_i = \exp(\mathbf{x}_i\mathbf{b})$, which yields the model known as NB2 with the mean function $b'(\theta) = \mu$ and variance function $b''(\theta) = \mu + \kappa\mu^2$. Thus, the Poisson model is included as a negative-binomial, i.e. $\kappa = 0$.

Interactions

Hosmer and Lemeshow (2000), Vittinghof *et al.* (2005), and others define the interaction as the changing slope parameter of one variable depending on the level of the other variable. Let us assume that we consider the interaction between variables x_1 and x_2 , which both may be categorical or continuous, and let the variable x_1 be a rating factor of interest. The changing slope parameter of x_1 depends on the level of x_2 , and thus

$$\beta_{x_1 \times x_2} x_1 = \beta_1 x_1 + \beta_{12} x_1 x_2 \text{ for all } x_2.$$

We can see that the coefficient $\beta_{x_1 \times x_2}$ varies according to the level of x_2 and can also have different signs for different values of x_2 , so the sign of $\beta_{x_1 \times x_2}$ is ambiguous. In addition, the statistical significance cannot be tested using the z-test (considering the coefficient β_{12}) because the standard error of the estimated changing parameter $\beta_{x_1 \times x_2}$ is in fact computed as

$$SE_{x_1 \times x_2} = \sqrt{x_1^2 \text{var}(\beta_1) + (x_1 x_2)^2 \text{var}(\beta_{12}) + 2x_1^2 x_2 \text{cov}(\beta_1; \beta_{12})}.$$

It follows that the z-statistics and the significance depend on the level of x_2 , which means that the interactions may be statistically significant only for some observations. Therefore, the likelihood ratio (LR) test based on deviance residuals is preferred to the z-test.

A very useful description of the interaction is the “treatment” effect representing the difference in outcome between the levels of x_2 , as predicted by the model. The treatment effect may be defined for the binary variable x_2 as $t(x_1) = \hat{g}_1(x_1) - \hat{g}_0(x_1)$.

Assessing the Fit of the Model

One of the residuals used to assess the fit of the model is Pearson residuals

$$r_i^p = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)}}$$

or deviance residual

$$r_i^d = \text{sign}(y_i - \mu_i) \sqrt{d(y_i, \mu_i)},$$

where *sign* is the function returning the sign of the argument and $d(y, \mu)$ is the distance function

$$d(y, \mu) = 2[yg(y) - b(g(y)) - yg(\mu) + b(g(\mu))]$$

which represents the distance of the estimated μ_i from y_i . Standardized Pearson or deviance residuals are divided by

$$\sqrt{\phi(1 - h_{ii})},$$

where h_{ii} is hat statistics which measure the influence of an observation on the outcome and is an element of the hat matrix

$$\mathbf{H} = \mathbf{V}^{1/2} \mathbf{X} (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{1/2},$$

where $\mathbf{X}$ is the design matrix and $\mathbf{V}$ is a diagonal matrix with general element $v_{ii} = (\partial \mu_i / \partial \mathbf{x}_i \mathbf{b})^2 V(\mu_i)^{-1}$.

Using deviance residuals, the deviance statistics is calculated to test the model by likelihood ratio test

$$\Delta D = D_0 - D_1,$$

where D_0 is the deviance of the null model and D_1 is the deviance of the estimated model and calculated as

$$D = D(\mathbf{y}, \boldsymbol{\mu}) = \frac{2}{\phi} \sum_{i=1}^N d(y_i, \mu_i).$$

It is also possible to employ the test to test models against each other. The LR test is approximately χ^2 -distributed with $p_0 - p_1$ degrees of freedom if the models have p_0 and p_1 parameters and if the condition $p_0 > p_1$ holds.

RESULTS AND DISCUSSION

In this section, we estimated the Poisson regression model and three negative-binomial models: (1) linear, (2) non-linear involving fractional polynomials and (3) the same non-linear model extended by interaction between given rating factors.

The data sample encompassed the characteristics of policies during the years 2004–2008 (220,022 observations). The following rating factors were included: vehicle age (*agecar*), engine volume (*volume*) and engine power (*kw*), owner's age (*ageman*), company car (*company*), gender of policyholder

(*gender*), type of fuel (*fuel*), price of the vehicle (*price*), and district area (*district*). We also knew the time duration of each policy (*duration*). In addition to the collinearity problem, a new variable (*kwvol*) that combines *volume* and *kw* was defined as follows: $kw/volume \cdot 1000$.

The non-linearity was identified for *ageman*, *agecar*, *kwvol*, *price* with FP powers of (2 2) for *kwvol*, (−2 3) for *ageman*, (.5 1) for *agecar*, and (0.5 0) for *price*. The interaction was considered only between *gender* × *ageman* because only this appeared to be reasonable and may be supported in practical terms.

Overdispersion

First, we compared the Poisson model and negative-binomial model to identify the overdispersion in the Poisson model. Next table compares the averaged predicted probability of both models.

According to the results, the Poisson model underestimated the probabilities of zero count and overestimated the occurrence of one insured accident. Contrary to the Poisson regression, the negative-binomial model fitted the observed probabilities better. In addition, the likelihood ratio test determining if the overdispersion parameter is statistically different from zero yielded a chi2, with one degree of freedom, 477.20. The corresponding

p-value was less than 0.001 which indicated that the negative binomial model $\kappa = 0.9912$ with was significant and the negative-binomial model should be preferred to the Poisson model. Furthermore, as shown in Fig. 1, the standardized deviance residuals of the negative-binomial model are significantly smaller.

Non-linear Effect of Rating Factors

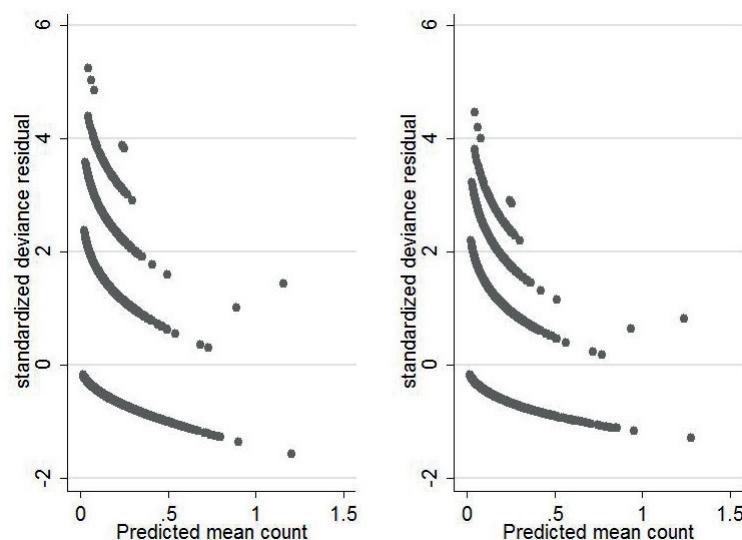
Next modelling issue affecting the accuracy and the quality of the model is transformation of variables. We estimated FPs of first and second degrees and tested them against each other and against the linear function using LR test. Tab. II summarizes the deviance difference and p-values corresponding to the chi2 distribution and given degrees of freedom.

Corresponding p-values in Tab. II indicated that the effect of rating factors is non-linear, thereby using FPs increased the fit of the model significantly. Thus, some technique involving modelling non-linearity should be considered in model-building process. The same conclusion can be drawn by comparing smoothed residuals, as shown in Fig. 2.

Clearly, the expected value of residuals corresponding to the linear function is significantly different from the zero (horizontal line at $y = 0$) which indicated the lack of fit. On the other hand,

I: Comparison of the Poisson and negative-binomial model: averaged predicted probability [%]

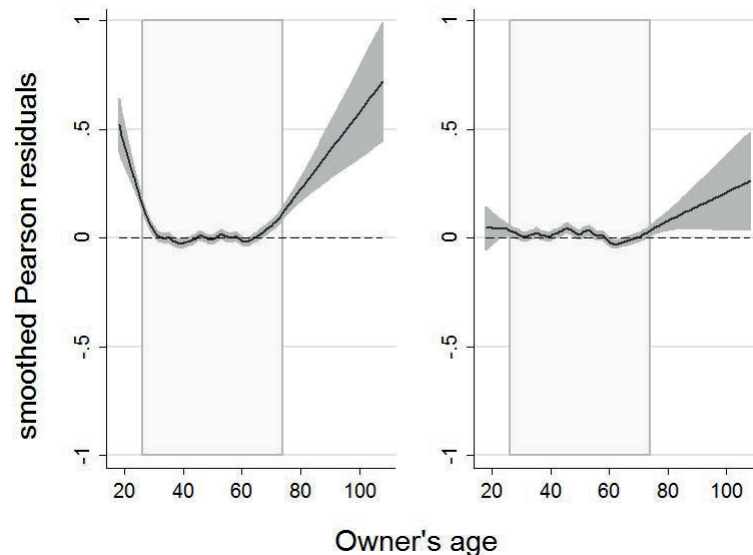
Count	Observed probability	Predicted probability		Difference	
		Poisson	Linear NB	Poisson	NB
0	93.562	93.348	93.603	0.214	−0.042
1	5.933	6.327	5.828	−0.393	0.105
2	0.459	0.312	0.505	0.147	−0.046
3	0.044	0.013	0.055	0.031	−0.011
4	0.002	0.001	0.007	0.002	−0.005



1: Comparison of the Poisson (left) and negative-binomial model (right): standardized deviance residuals

II: Comparison of non-linear and linear effects of given rating factors

Variable	p-values		Deviance difference	
	FP2 against linear	FP2 against FP1	FP2 against linear	FP2 against FP1
agecar	0.000	0.000	44.357	19.664
ageman	0.000	0.000	274.553	21.631
kwvol	0.001	0.008	17.494	9.782
price	0.000	0.000	36.904	36.904



2: Comparison of smoothed Pearson residuals with 95% confidence interval plotted as a shaded area: the linear (left) and fractional polynomial (right) function

the expected residuals corresponding to FP lies around the zero in 95% confidence interval. The better fit of extremal observations is also obvious (the lightly shaded box within the plot region represents an interval in which the 95% of the observations lie).

Thus, it is obvious that involving FPs fitted the data better. However, to model non-linear effects using this method, continuous rating factors are necessary and the computation is generally time-consuming. In addition, to ensure the transferability of the model, the stability analysis of FPs should be undertaken and also sufficient validation of the model should be considered.

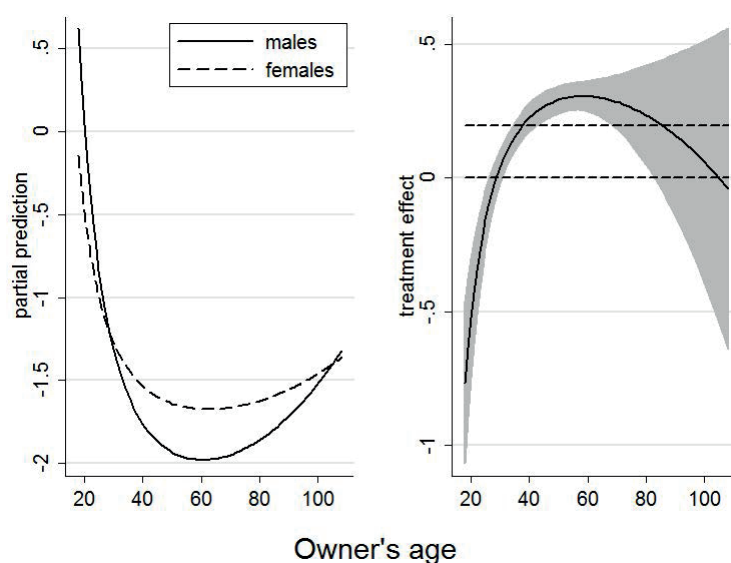
Interacted Rating Factors

The last issue to be addressed in this study is modelling interactions. We represented here the interaction between *ageman* $\times$ *gender* which was statistically significant and can be also explained in rational terms. Thus, by adding this interaction in the claim frequency model, we extended the NB2 model involving fractional polynomials. The LR test determining the statistically significant difference from the NB2 model without interaction term yielded the χ^2 , with 2 degrees of freedom, 43.45 and the corresponding p-value was less than 0.0001 which indicated better fit of the model. In addition,

we estimated the partial prediction of *ageman* dependently on the *gender* and also the “treatment” effect with 95% confidence level (the shaded area), as shown in Fig. 3.

Clearly, the functions that represent the effects of owner's age differed according to the gender. By focusing on this treatment effect plot, we can conclude that the effects of gender also differed significantly from the main effect, but the effect also changed from negative to positive. By comparing these effects, the interaction implies that claims are less likely from young women than young men in the 18–24 years interval, whereas they are more likely over 24 years whereas the claim is more likely for females for any age if the model without interaction is considered.

Thus, modelling interaction may also increase the fit of the model as well as modelling non-linearity. Searching for an interaction might yield a model with a better fit, or determine a potential interaction to be analysed further. However, the power of interaction test is weak at detecting even moderately large interactions, mainly due to small sample sizes. In addition, the power of the test is greater if the factor is continuous, rather than binary or categorical, Farewell, Tom and Royston (2004). Thus, it follows that a large sample size encompassing continuous rating factors is an unconditional



3: Effects of owner's age on gender. Left: the effect of owner's age on claim frequency. Right: the "treatment" effect (solid line) and the main effect of gender (dashed line)

prerequisite for any reasonable interaction analysis. However, the computation of interactions is not as time-demanding as using FPs, but the sceptical approach should generally be applied when

deciding which interactions to include in the model because each interaction should be reasonably interpreted.

CONCLUSION

The paper was focused on modelling claim frequency in vehicle insurance. We showed that the overdispersion, non-linear systematic component and interacted rating factors should be considered when the claim frequency is modelled.

First, we detected overdispersion in the Poisson model and employed the negative-binomial model to show that considering heterogeneity in insurance policies yields better fit of the model, as well as more precise estimates of claim frequency. Second, we analysed the linear effect of continuous rating factors. We showed that using non-linear systematic component also yield the better fit of the model, as well as generate new hypotheses to be verified further, especially the stability of the fractional polynomial functions. Finally, using interaction between owner's age and gender, we demonstrated that considering interactions in the model may also increase the fit of the model and generate new hypotheses to be analysed further, i.e. the findings that the effect of gender is changing dependently on the owner's age.

Thus, we concluded that considering overdispersion is a crucial point in modelling claim frequency. Using fractional polynomials and interactions are disputable because a large data sample is necessary. On the one hand, such dataset are available to insurance companies, on the other hand, these modelling techniques are very computationally demanding. In addition, the transferability of the model must be considered and therefore sufficient validation of the model is necessary.

Acknowledgement

This paper was written within the project SP Application of generalized linear models to insurance and finance – Project No. 2015/75 and within the Operational Programme Education for Competitiveness – Project No. CZ.1.07/2.3.00/20.0296.

REFERENCES

- ALTMAN, D. G., LAUSEN, B. SAUERBREI, W. et al. 1994. The danger of using 'optimal' cutpoint in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86: 829–835.
- ANDRADE-SILVA, J. M. 1989. An application of Generalized Linear Models to Portuguese Motor Insurance. In: *Proceedings XXI ASTIN Colloquium*. New York, 633.
- BRANDA, M. 2014. Optimization Approaches to Multiplicative Tariff of Rates Estimation in Non-Life Insurance. *Asia-Pacific Journal of Operational Research*, 31(5): 1450032.

- BROCKMAN, M. J. and WRIGHT, T. S. 1992. Statistical motor rating: making efficient use of your data. *Journal of the Institute of Actuaries*, 119(3): 457–543.
- FAREWELL, V. T., TOM, B. D. M. and ROYSTON, P. 2004. The impact of dichotomization on the efficiency of testing for an interaction effect in exponential family models. *Journal of the American Statistical Association*, 99: 822–831.
- HOSMER, D. W. and LEMESHOW, S. 2000. *Applied Logistic Regression*. 2nd ed. John Wiley & Sons.
- JONG, P. de and HELLER, G. Z. 2008. *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press.
- KAFKOVÁ, S. and KŘIVÁNKOVÁ, L. 2014. Generalized Linear Models in Vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62(2): 383–388.
- KAHANE, Y. and LEVY, H. 1975. Regulation in the insurance industry: Determination of premiums in automobile insurance. *Journal of Risk and Insurance*, 42: 117–132.
- MARCUS, R. PERITZ, E. and GABRIEL, K. R. 1976. On closed test procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3): 655–660.
- MCCULLAGH, P. and NELDER, J. A. 1983. *Generalized linear models*. London: Chapman & Hall.
- MCCULLAGH, P. and NELDER, J. A. 1989. *Generalized linear models*. 2nd ed. London: Chapman & Hall.
- MURPHY, K. P., BROCKMAN, M. J. and LEE, P. K. W. 2000. Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance. In: *Casualty Actuarial Society Forum*. 107–140.
- OHLSSON, E. and JOHANSSON, B. 2010. *Non-Life Insurance Pricing with Generalized Linear Models*. Berlin: Springer.
- RENSHAW, A. E. 1994. Modelling the claims process in the presence of covariates. *ASTIN Bulletin*, 24(2): 265–285.
- ROYSTON, P. and ALTMAN, D. G. 1994. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modeling. *Applied Statistics*, 43(3): 429–467.
- VITTINGHOF, E. GLIDDEN, D. V., SHIBOSKI, S. C. et al. 2005. *Regression Methods in Biostatistics. Linear, Logistic, survival, and Repeated Measures Models*. New York: Springer.
- WOOD, S. N. 2006. *Generalized Additive Models*. Boca Raton: Chapman & Hall/CRC.
- ZAKS, Y., FROSTIG, E. and LEVIKSON, B. 2006. Optimal pricing of a heterogeneous portfolio for a given risk level. *ASTIN Bulletin*, 36(1): 161–185.