# Underwriting risk control in non-life insurance via generalized linear models and stochastic programming

Martin Branda [1]

**Abstract.** We focus on rating of non-life insurance contracts. We employ multiplicative models with basic premium levels and specific surcharge coefficients for various levels of selected segmentation criteria (rating factors). We use generalized linear models to describe the probability distribution of total losses for a contract during one year. In particular, overdispersed Poisson regression is used to model the expected number of claims during a given period and Gamma or Inverse-Gaussian regression are applied to predict average claim severity. We propose stochastic programming problems with reliability type constraints for the surcharge coefficients estimation which take into account riskiness of each rate cell, prescribed loss ratio and other business requirements. We apply the approach to Motor Third Party Liability (MTPL) policies.

**Keywords:** non-life insurance, rate making, generalized linear models, stochastic programming, MTPL.

**JEL classification:** C44
**AMS classification:** 90C15

## 1 Introduction

Traditional credibility models in non-life insurance take into account known history of a policyholder and project it into policy rate, see [8]. However, for new business, i.e. new clients coming for an insurance policy, the history need not to be known or the information may not be reliable. Thus traditional approaches to credibility can not be used. We will employ models which are based on settled claims of new contracts from the previous years. This experience is transfered using generalized linear models (GLM), see [14], which cover many important regression models and are nowadays widely applied in insurance, cf. [1, 9, 12, 15]. Expected claim count on a policy during one year and expected claim size can be explained by various independent variables which can serve as segmentation criteria, e.g. age and gender of the policyholder and region where he or she lives. Using these criteria and GLM we can derive surcharges which enable to take into account riskiness of each policy. However, as we will show in this paper, these coefficients need not to fulfill business requirements, for example restriction on maximal surcharge. Optimization models must be then employed.

Stochastic programming techniques can be used to solve optimization problems where random coefficients appear. In this paper, we will employ formulation based on reliability type constraints such as chance constraints and the reformulation based on one-sided Chebyshev inequality. The distribution of the random parts will be represented by compound Gamma-Poisson and Inverse Gaussian-Poisson distributions with parameter estimates based on generalized linear models. Sensitivity analysis of the results with respect to the underlying distribution is often proposed, cf. [2, 6, 7, 10, 13].

This paper is organized as follows. In Section 2, we will review definition and basic properties of generalized linear models. Rate-making approach based on GLM is then proposed in Section 3. In Section 4, optimization models for rates estimation are introduced which enable to take into account various business requirements. These models are extended using stochastic programming techniques in Section 5. Section 6 concludes the paper with an application of the proposed methods to MTPL contracts.

[1]Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Praha 8, branda@karlin.mff.cuni.cz

## 2 Generalized linear models

In this section, we introduce generalized linear models (GLM) which cover many useful regression models. GLM are based on the following three building blocks:
**1.** The dependent variable $Y_i$ has distribution from the exponential family with probability density function

$$f(y;\theta_i,\varphi) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\varphi} + c(y,\varphi)\right\}, \tag{1}$$

where $b, c$ are known functions and $\theta_i, \varphi$ are unknown canonical and dispersion parameters.
**2.** A linear combination of independent variables is considered

$$\eta_i = \sum_j X_{ij}\beta_j, \tag{2}$$

where $\beta_j$ are unknown parameters and $X_{ij}$ are given values of predictors.
**3.** The dependency is described by a link function $g$ which is strictly monotonous and twice differentiable

$$\mathbb{E}[Y_i] = \mu_i = g^{-1}(\eta_i). \tag{3}$$

The most important members of the exponential family are proposed in Table 1 including basic characteristics which are introduced below. The following relations can be obtained for expectation and variance under the assumption that $b$ is twice continuously differentiable

$$\mathbb{E}[Y] = b'(\theta), \tag{4}$$
$$var(Y) = \varphi b''(\theta) = \varphi V(\mu), \tag{5}$$

where the last expression is rewritten using the variance function which is defined as $V(\mu) = b''[(b')^{-1}(\mu)]$.

| **Distribution** | **Density** $f(y;\theta,\varphi)$ | **Dispersion param.** $\varphi$ | **Canonical param.** $\theta(\mu)$ | **Mean value** $\mu(\theta)$ | **Variance function** $V(\mu)$ |
|---|---|---|---|---|---|
| $N(\mu,\sigma^2)$ | $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ | $\sigma^2$ | $\mu$ | $\theta$ | 1 |
| $Po(\mu)$ | $\frac{\mu^y e^{-\mu}}{y!}$ | 1 | $\log(\mu)$ | $e^\theta$ | $\mu$ |
| $\Gamma(\mu,\nu)$ | $\frac{1}{\Gamma(\nu)y}\left(\frac{y\nu}{\mu}\right)^\nu e^{-\frac{y\nu}{\mu}}$ | $\frac{1}{\nu}$ | $-\frac{1}{\mu}$ | $-\frac{1}{\theta}$ | $\mu^2$ |
| $IG(\mu,\lambda)$ | $\sqrt{\frac{\lambda}{2\pi y^3}}e^{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}}$ | $\frac{1}{\lambda}$ | $-\frac{1}{2\mu^2}$ | $\frac{1}{\sqrt{-2\theta}}$ | $\mu^3$ |

**Table 1** Distributions from the exponential family

Maximum likelihood method is used to estimate the parameters of GLM. For overdispersed Poisson model where the variance need not to be equal to the expected value (dispersion $\varphi$ is not set to 1 but is estimated from data) quasi-likelihood function must be used. For details see [14].

## 3 Rate-making using generalized linear models

We denote $i_0 \in \mathcal{I}_0$ the basic segmentation criterion, e.g. tariff groups, and $i_1 \in \mathcal{I}_1, \ldots, i_S \in \mathcal{I}_S$ the other segmentation criteria which should help us to take into account underwriting risk. We will denote one risk cell $I = (i_0, i_1, \ldots, i_S)$ with $I \in \mathcal{I} = \mathcal{I}_0 \otimes \mathcal{I}_1 \otimes \cdots \otimes \mathcal{I}_S$. Let $L_I = \sum_{n=1}^{N_I} X_{In}$ denote aggregated losses over one year for risk cell $I$ where $N_I$ is the random number of claims and $X_{In}$ the random claim severity. All the variable are assumed to be independent. Then, for the mean and the variance it holds

$$\mu_I = \mathbb{E}[L_I] = \mathbb{E}[N_I]\mathbb{E}[X_I], \tag{6}$$
$$\sigma_I^2 = var(L_I) = \mathbb{E}[N_I]var(X_I) + (\mathbb{E}[X_I])^2 var(N_I). \tag{7}$$

The premium is based on multiplicative model composed from basic premium levels $Pr_{i_0}$ and non-negative surcharge coefficients (rating factors) $e_{i_1}, \ldots, e_{i_S}$, i.e.

$$Pr_I = Pr_{i_0} \cdot (1 + e_{i_1}) \cdot \cdots \cdot (1 + e_{i_S}). \tag{8}$$

Our goal is to find the optimal basic premium levels and coefficient with respect to a prescribed loss ratio $\hat{LR}$, i.e. we would like to fulfill the constraints

$$\frac{L_I}{Pr_I} \leq \hat{LR} \text{ for all } I \in \mathcal{I}. \tag{9}$$

The goal loss ratio $\hat{LR}$ is usually based on management decision. It is possible to prescribe different loss ratios for each tariff cell but this is not considered in this paper. However, losses $L_I$ are random. The simplest way is to hedge against expected value of losses $\mathbb{E}[L_I]$. This can be done directly using GLM with logarithmic link function.

Poisson distribution and Gamma or Inverse Gaussian without intercept are used to estimate parameters for expected number of claims and severity. If we use logarithmic link function in both regression models, then we can get for $I = (i_0, i_1, \ldots, i_S)$

$$\mathbb{E}[N_I] = \exp\{\lambda_{i_0} + \lambda_{i_1} + \cdots + \lambda_{i_S}\}, \tag{10}$$
$$\mathbb{E}[X_I] = \exp\{\gamma_{i_0} + \gamma_{i_1} + \cdots + \gamma_{i_S}\}, \tag{11}$$

where $\lambda_i, \gamma_i$ are estimated coefficients. Thus for the mean loss it holds

$$\mathbb{E}[L_I] = \exp\{\lambda_{i_0} + \gamma_{i_0} + \lambda_{i_1} + \gamma_{i_1} + \cdots + \lambda_{i_S} + \gamma_{i_S}\}. \tag{12}$$

Now, if we set $\hat{\lambda}_i = \exp(\lambda_i)/min_{i\in I}\exp(\lambda_i)$ and $\hat{\gamma}_i = \exp(\gamma_i)/min_{i\in I}\exp(\gamma_i)$, the basic premium levels and surcharge coefficient can be estimated as

$$Pr_{i_0} = \frac{\exp\{\lambda_{i_0} + \gamma_{i_0}\}}{\hat{LR}} \prod_{s=1}^{S} \min_{i\in\mathcal{I}_s} \exp(\lambda_i) \prod_{s=1}^{S} \min_{i\in\mathcal{I}_s} \exp(\gamma_i), \tag{13}$$
$$1 + e_{i_s} = \exp\{\lambda_{i_s} + \gamma_{i_s}\}, \tag{14}$$

Then the constraints (9) are fulfilled in expectation. However, the surcharge coefficient estimates often violate business requirements, especially they can be too high, as we will show in the numerical study.

## 4 Optimization problem for rate estimation

The constraints (9) with expectation can be rewritten as

$$\mathbb{E}[L_{i_0,i_1,\ldots,i_S}] \leq \hat{LR} \cdot Pr_{i_0} \cdot (1 + e_{i_1}) \cdot \cdots \cdot (1 + e_{i_S}). \tag{15}$$

There can be set business limitation that the highest aggregated risk surcharge is lower than a given level $r^{max}$. We would to minimize basic premium levels and surcharges which are necessary to fulfill the prescribed loss ratio and the business requirements. This leads to the following nonlinear optimization problem

$$\begin{aligned} \min \prod_{i_0\in\mathcal{I}_0} Pr_{i_0} \prod_{i_1\in\mathcal{I}_1} (1 + e_{i_1}) \cdot \cdots \prod_{i_S\in\mathcal{I}_S} (1 + e_{i_S}) & \\ \text{s.t.} & \\ \hat{LR} \cdot Pr_{i_0} \cdot (1 + e_{i_1}) \cdot \cdots \cdot (1 + e_{i_S}) &\geq \mathbb{E}[L_{i_0,i_1,\ldots,i_S}],\ (i_0, i_1, \ldots, i_S) \in \mathcal{I}, \\ (1 + e_{i_1}) \cdot \cdots \cdot (1 + e_{i_S}) &\leq 1 + r^{max}, \\ e_{i_1}, \ldots, e_{i_S} &\geq 0. \end{aligned} \tag{16}$$

Using logarithmic transform of the decision variables $u_{i_0} = \ln(Pr_{i_0})$ and $u_{i_s} = \ln(1 + e_{i_s})$ and by setting $b_{i_0,i_1,\ldots,i_S} = \ln(\mathbb{E}[L_{i_0,i_1,\ldots,i_S}]/\hat{LR})$ the problem can be rewritten as linear programming problem which can be efficiently solved by standard software tools.

$$\begin{aligned} \min \sum_{i_0\in\mathcal{I}_0} u_{i_0} + \sum_{i_1\in\mathcal{I}_1} u_{i_1} + \cdots + \sum_{i_S\in\mathcal{I}_S} u_{i_S} & \\ \text{s.t.} & \\ u_{i_0} + u_{i_1} + \cdots + u_{i_S} &\geq b_{i_0,i_1,\ldots,i_S},\ (i_0, i_1, \ldots, i_S) \in \mathcal{I}, \\ u_{i_1} + \cdots + u_{i_S} &\leq \ln(1 + r^{max}), \\ u_{i_1}, \ldots, u_{i_S} &\geq 0. \end{aligned} \tag{17}$$

| | | **Overd. Poisson** | | | **Gamma** | | | **Inv. Gaussian** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Param. | Level | Est. | Std.Err. | Exp | Est. | Std.Err. | Exp | Est. | Std.Err. | Exp |
| tariff group | 1 | -3.096 | 0.042 | 0.045 | 10.30 | 0.015 | 29 778 | 10.30 | 0.017 | 29 765 |
| tariff group | 2 | -3.072 | 0.038 | 0.046 | 10.35 | 0.013 | 31 357 | 10.35 | 0.015 | 31 380 |
| tariff group | 3 | -2.999 | 0.037 | 0.050 | 10.46 | 0.013 | 34 913 | 10.46 | 0.015 | 34 928 |
| tariff group | 4 | -2.922 | 0.037 | 0.054 | 10.54 | 0.013 | 37 801 | 10.54 | 0.015 | 37 814 |
| tariff group | 5 | -2.785 | 0.040 | 0.062 | 10.71 | 0.014 | 44 666 | 10.71 | 0.017 | 44 679 |
| region | 1 | 0.579 | 0.033 | 1.785 | 0.21 | 0.014 | 1.234 | 0.21 | 0.016 | 1.234 |
| region | 2 | 0.460 | 0.031 | 1.583 | 0.11 | 0.013 | 1.121 | 0.11 | 0.014 | 1.121 |
| region | 3 | 0.205 | 0.032 | 1.228 | 0.06 | 0.013 | 1.059 | 0.06 | 0.015 | 1.058 |
| region | 4 | 0.000 | 0.000 | 1.000 | 0.00 | 0.000 | 1.000 | 0.00 | 0.000 | 1.000 |
| age | 1 | 0.431 | 0.027 | 1.539 | - | - | - | - | - | - |
| age | 2 | 0.245 | 0.024 | 1.277 | - | - | - | - | - | - |
| age | 3 | 0.000 | 0.000 | 1.000 | - | - | - | - | - | - |
| gender | 1 | -0.177 | 0.018 | 0.838 | - | - | - | - | - | - |
| gender | 2 | 0.000 | 0.000 | 1.000 | - | - | - | - | - | - |
| Scale | | 0.647 | 0.000 | | 13.84 | 0.273 | | 0.002 | 0.000 | |

**Table 2** Parameter estimates of GLM

## 5 Stochastic programming problems for rate estimation

In this section, we propose stochastic programming formulations which take into account compound distribution of random losses not only its expected value. We employ chance constraints for satisfying the constraints (9). However, chance constrained problems are very computationally demanding in general, see [3, 4, 5, 6, 11] for various solution approaches and possible reformulations.

If we prescribe a probability level $\varepsilon$ for violating the prescribed loss ratio in each tariff cell, we obtain the following chance (probabilistic) constraints

$$P(L_{i_0,i_1,\ldots,i_S} \leq \hat{LR} \cdot Pr_{i_0} \cdot (1+e_{i_1}) \cdot \cdots \cdot (1+e_{i_S})) \geq 1-\varepsilon, \tag{18}$$

which can be rewritten using quantile function of $L_{i_0,i_1,\ldots,i_S}$ as

$$\hat{LR} \cdot Pr_{i_0} \cdot (1+e_{i_1}) \cdot \cdots \cdot (1+e_{i_S}) \geq F^{-1}_{L_{i_0,i_1,\ldots,i_S}}(1-\varepsilon) \tag{19}$$

Setting $b_{i_0,i_1,\ldots,i_S} = \ln[F^{-1}_{L_{i_0,i_1,\ldots,i_S}}(1-\varepsilon)/\hat{LR}]$ formulation (17) can be used. However, it can be very difficult to compute the quantiles for compound distribution, see [16]. Instead of approximating the quantiles, we can employ one-sided Chebyshev inequality based on the mean and variance of the the compound distribution resulting in the constraints

$$P(L_I \geq Pr_I) \leq \frac{1}{1+(Pr_I-\mu_I)^2/\sigma_I^2} \leq \varepsilon, \text{ for } Pr_I \geq \mu_I, \tag{20}$$

which can be rewritten as

$$\frac{1-\varepsilon}{\varepsilon}\sigma_I^2 \leq (Pr_I-\mu_I)^2. \tag{21}$$

This leads to the following reliability constraints

$$\mu_I + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\sigma_I \leq Pr_I. \tag{22}$$

Setting $b_I = \ln[(\mu_I + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\sigma_I)/\hat{LR}]$ we can employ linear programming formulation (17).

| | | **GLM** | | **MV-model** | | **SP-model I** | | **SP-model II** | |
|---|---|---|---|---|---|---|---|---|---|
| | Level | G | IG | G | IG | G | IG | G | IG |
| tariff group | 1 | 1 880 | 1 879 | 3 881 | 3 877 | 127 164 | 253 227 | 7 916 | 12 112 |
| tariff group | 2 | 2 028 | 2 029 | 4 186 | 4 187 | 135 565 | 277 303 | 8 483 | 13 209 |
| tariff group | 3 | 2 430 | 2 431 | 5 017 | 5 016 | 156 748 | 337 577 | 9 976 | 16 003 |
| tariff group | 4 | 2 840 | 2 841 | 5 863 | 5 862 | 176 535 | 394 915 | 11 437 | 18 715 |
| tariff group | 5 | 3 850 | 3 851 | 7 948 | 7 946 | 223 966 | 542 834 | 14 993 | 25 627 |
| region | 1 | 2.203 | 2.201 | .241 | .240 | .464 | .470 | .293 | .358 |
| region | 2 | .775 | .776 | .000 | .000 | .250 | .200 | .077 | .105 |
| region | 3 | .301 | .299 | .000 | .000 | .037 | .000 | .000 | .000 |
| region | 4 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| age | 1 | .539 | .539 | .350 | .351 | .248 | .244 | .363 | .316 |
| age | 2 | .277 | .277 | .121 | .121 | .133 | .132 | .188 | .166 |
| age | 3 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| gender | 1 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| gender | 2 | .194 | .194 | .194 | .194 | .095 | .094 | .135 | .119 |

**Table 3** Optimal rates and segmentation coefficient

## 6 Numerical example

In this section, we apply proposed approaches to Motor Third Party Liability contracts. We consider 60000 policies which are simulated using characteristics of real MTPL portfolio of one of the leading Czech insurance companies. The following criteria are used in GLM as the independent variables:
**1. tariff group**: 5 categories (up to 1000, over 1350, over 1850, over 2500, over 2500 ccm engine),
**2. age**: 3 categories (18-30, 30-65, 65 and more years),
**3. region**: 4 categories (over 500 000, over 50 000, over 5 000, up to 5 000 inhabitants),
**4. gender**: 2 categories (men, women).

We employ the approaches proposed in the previous sections to find the basic premium levels for the tariff groups and the surcharge coefficients for other criteria. The goal loss ratio for new business is set to 0.6 and the maximum feasible surcharge to 100 percent. The parameter estimates for overdipersed Poisson, Gamma (G) and Inverse Gaussian (IG) generalized linear models can be found in Table 2. Standard errors and exponentials of the coefficient are also included. All variables are significant based on Wald and likelihood-ratio tests. The parameters of GLM were estimated using SAS GENMOD procedure [17] and the optimization problems were solved using SAS OPTMODEL procedure [18].

The basic premium levels and surcharge coefficients can be found in Table 3. It is not surprising that the coefficients which are estimated directly from GLM do not fulfill the business requirements and the highest possible surcharge is much higher than 100 percent. This drawback is removed by the optimization problems. The decrease of the surcharge coefficient leads to the increase of the basic premium levels. We refer to the problem where the expected loss is covered as MV-model. Inappropriate increase of rates can be observed if we use directly the stochastic programming formulation with the reliability type constraints based on Chebyshev inequality with $\varepsilon = 0.1$ (SP-model I). This increase is reduced in the second stochastic programming problem (SP-model II), where lower "weights" (0.1) are assigned to the variance in formula (22). Stochastic programming models with Inverse Gaussian regression lead to higher estimates of the basic premium levels because the estimated variance is much higher then for Gamma regression. Thus, the first model leads to safer estimates however the variance observed in practice corresponds rather to Gamma model.

## 7 Conclusion

In this paper, we compared several methods for rating of non-life (MTPL) insurance contracts which take into account riskiness of various segments. The probability distribution of losses was described by generalized linear models. Direct application of the estimated coefficient leads to the surcharge coeffi-

cients which do not fulfill the business requirements. Therefore, optimization models were introduced. Stochastic programming formulation was employed to consider the distribution of the random losses on a policy.

## Acknowledgements

This work was supported by the Czech Science Foundation under grant [P402/12/G097].

## References

[1] Antonio, K., and Beirlant, J.: Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics* **40** (2007), 58–76.

[2] Branda, M.: Local stability and differentiability of the Conditional Value at Risk defined on the mixed-integer loss functions. *Kybernetika* **46**(3) (2010), 362–373.

[3] Branda, M.: Stochastic programming problems with generalized integrated chance constraints, Accepted to *Optimization*, 2012. DOI: 10.1080/02331934.2011.587007

[4] Branda, M.: Chance constrained problems: penalty reformulation and performance of sample approximation technique. *Kybernetika* **48**(1) (2012), 105–122.

[5] Branda, M.: Sample approximation technique for mixed-integer stochastic programming problems with several chance constraints. *Operations Research Letters* **40**(3) (2012), 207–211.

[6] Branda, M., and Dupačová, J.: Approximations and contamination bounds for probabilistic programs. *Annals of Operations Research* **193**(1) (2012), 3–19.

[7] Branda, M., and Kopa, M.: DEA-Risk Efficiency and Stochastic Dominance Efficiency of Stock Indices. *Czech Journal of Economics and Finance* (Finance a uver) **62**(2) (2012), 106-124.

[8] Bühlmann, H., and Gisler, A.: *A course in credibility theory and its applications*. Springer Science & Business, 2005.

[9] Denuit, M., Marchal, X., Pitrebois, S., Walhin, J.-F.: *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons, Chichester, 2007.

[10] Dupačová, J., and Kopa, M.: Robustness in stochastic programs with risk constraints. *Annals of Operations Research* (Online first), 2012. DOI: 10.1007/s10479-010-0824-9.

[11] Houda, M., and Kaňková, V.: Empirical Estimates in Economic and Financial Optimization Problems. *Bulletin of the Czech Econometric Society* **19**(29) (2012), 50–69.

[12] de Jong, P., and Heller, G.Z.: *Generalized Linear Models for Insurance Data*. Cambridge University Press, New York, 2008.

[13] Lachout, P.: Approximative solutions of stochastic optimization problems. *Kybernetika* **46**(3) (2010), 513–523.

[14] McCullagh, P., and Nelder, J.A.: *Generalized Linear Models*. 2nd Ed. Chapman and Hall, London, 1989.

[15] Ohlsson, B.: Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal* **4** (2008), 301–314

[16] Withers, Ch., and Nadarajah, S.: On the compound Poisson-gamma distribution. *Kybernetika* **47**(1) (2011), 15–37.

[17] SAS/STAT 9.3: User's Guide.

[18] SAS/OR 9.3 User's Guide: Mathematical Programming.