# MODELLING OF INDIVIDUAL INSURED ACCIDENT RISK FOR GIVEN MOTOR-HULL INSURANCE PORTFOLIO

Jiří Valecký

**Abstract**

Modelling of individual insured accident risk is important in terms of insurer's business policy and risk managing. Firstly, the insurance company may want to differentiate the insurance premium in order to determine more fair insurance price and to gain new policyholders. Secondly, the insured accident probability is used in order to estimate the insurance risk capital requirement within the SOLVENCY II. The paper is devoted to developing of a preliminary statistical count model appropriate to the modelling of insured accident probability for given motor-hull insurance portfolio in the Czech Republic. Firstly, the Poisson model is considered and because of the fact that the data violates the variance assumption, we turn to the negative-binomial model. To deal with the misspecified link function, we employ the fractional polynomials and finally, all models are compared by some statistical tests and by residual analysis.

**Key words:** insured accident, count models, fractional polynomial, motor-hull insurance, negative-binomial model

**JEL Code:** C25, C58, G22

## Introduction

Insurance premium rate is determined per monetary unit in accordance with undertaken risk. It means that the premium rate corresponds to the probability of insured accident and more risky client pays more. This trend has been already observed for many years already. It is very common that the insurers set the premium in motor-hull insurance in compliance with the volume of an engine. Nowadays, some of them also distinguish the size of district where the client lives and some insurers respect even client´s age.

To evaluate the undertaken risk represented by of insured accident probability, the models from GLM family, firstly formulated by (Nelder & Wedderburn, 1972), are applied in this area and one can distinguish two types of the model employed here, binomial and count models. Whereas the binomial models are primarily focused on the probability of success or failure, the latter are focused on the prediction of the count of successes or failures. The count probability is mostly determined in accordance with Poisson distribution. However the Poisson model is rarely sufficient to use and it is taken as a benchmark mostly. The real-data are mostly overdispersed with conditional variance exceeding conditional mean which violates the assumption of Poisson distributions. Therefore, the variance function was modified by adding an unobserved heterogeneity factor, firstly parameterized by (Bliss & Owen, 1958) and favored by (Breslow, 1984) and (Lawless, 1987).

Another problem of the general regression models is misspecified link function which may be tested by link test, based on an idea of (Tukey, 1949) and further described by (Pregibon, 1980). It is true that another link function may be used and regardless to the fact that the link function is often in non-canonical form, the link function may be misspecified due to the neglecting the fact that risk factor affects the outcome differently for various values and thus the relation between them is nonlinear. Here, the fractional polynomials (FPs) may be very helpful, see (Royston & Altman, 1994) or (Royston, Ambler, & Sauerbrei, 1999).

The aim of this paper is to propose the preliminary count model appropriate to the modelling of insured accident probability for given motor-hull insurance portfolio. We focus on the well fitted model and therefore the Poisson model is firstly considered as a benchmark only. Then, we turn to the negative-binomial model due to the overdispersion and to deal with the misspecified link function, we employ the fractional polynomials. Finally, we compare all models by conducting of some statistical tests and by residual analysis. The paper is organized as follows. In Section 1, there is derived a non-linear negative-binomial model. The proposed count models are estimated and compared in Section 2. The last section concludes the paper.

## 1 Derivation of nonlinear negative-binomial model

Here, we derive the general negative-binomial model known as NB2. In spite of the fact that this model has non-canonical link function, we use it because this modification of NB is generally mostly applied, see (Hilbe, 2011, s. 208).

### 1.1 Derivation of negative-binomial model

The Negative-binomial model (NB2) is usually derived from the Poisson model with gamma heterogeneity where the gamma noise has a mean of 1. Let's write a distribution function in the form of

$$f\left(y;\alpha,\mu\right)=\frac{\exp\left(-\alpha_i\mu_i\right)\left(\alpha_i\mu_i\right)^{y_i}}{y_i\,!}, \tag{1}$$

where $\mu_i$ is an individual intensity $\exp(x_i\beta)$ and $\alpha_i$ is a random heterogeneity factor $\exp(\varepsilon_i)$.

From the Equation (1) we obtain the negative binomial p.d.f., see (Hardin & Hilbe, 2007) for more details,

$$f(y;\mu,\alpha)=\frac{\Gamma(y_i+1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(1-\frac{1}{1+\alpha\mu_i}\right)^{y_i} \tag{2}$$

and general log-likelihood function

$$\ell(\mu;y,\alpha)=\sum_{i=1}^{n}\left\{\begin{aligned}&y_i\ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)-\left(\frac{1}{\alpha}\right)\ln(1+\alpha\mu_i)+\dots\\&\dots+\ln\Gamma\left(y_i+\frac{1}{\alpha}\right)-\ln\Gamma(y_i+1)-\ln\Gamma\left(\frac{1}{\alpha}\right)\end{aligned}\right\}. \tag{3}$$

Inserting the inverse link function in the form of $\mu_i=\exp(x_i\beta)$ into (3), we obtain the log-likelihood function

$$\ell(\mu;y,\alpha)=\sum_{i=1}^{n}\left\{\begin{aligned}&y_i\ln\left(\frac{\alpha\exp(x_i\beta)}{1+\alpha\exp(x_i\beta)}\right)-\left(\frac{1}{\alpha}\right)\ln\left(1+\alpha\exp(x_i\beta)\right)+\\&+\ln\Gamma\left(y_i+\frac{1}{\alpha}\right)-\ln\Gamma(y_i+1)-\ln\Gamma\left(\frac{1}{\alpha}\right)\end{aligned}\right\} \tag{4}$$

where the parameters $\alpha,\beta$ are estimated using the Newton-Raphson or IRLS algorithm.

## 1.2 Fractional polynomials

The expression $x_i\beta=g(x)$ in log-likelihood function (4) may not necessary be linear. One of the techniques how to handle with nonlinearity is using fractional polynomials (FPs). Thus, let's define the nonlinear function and rewrite the expression $g(x)$ into the form of

$$g(x)=\beta_0+\sum_{j=1}^{J}\beta_{1j}F_j(x_1)+\beta_2x_2+\dots+\beta_Kx_K, \tag{5}$$

where $F_j(x_1)$ is a particular type of power function. The power $p_j$ could be any number, but (Royston & Altman, 1994) restricts the power to be among 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

$$F_j(x_1)=\begin{cases}x_1^{p_j},\, p_j\neq p_{j-1},\\F_{j-1}(x_1)\ln(x_1),\, p_j=p_{j-1}.\end{cases} \tag{6}$$

The identification and comparison of the most appropriate FPs is made by closed test procedure, (Marcus, Peritz, & Gabriel, 1976), which is generally preferred over the sequential procedure, (Royston & Altman, 1994).

## 2 Empirical count models for given insurance portfolio

In this section, we propose the preliminary count model which could be used for the purpose of modelling insured accident probability. Here, we consider the most applied, although non-canonical, negative-binomial model known as NB2 model described above. All models are estimated on a data sample encompassing characteristics of policyholders in motor-hull insurance portfolio during the year 2008 (61 900 of insurance policies). In our study, we consider several variables, i.e. age of a car (*agecar*) volume (*volume*) and performance of the engine (*kw*), age of a policyholder (*ageman*), number of citizens in a region (*nocit*), company car (*company*), gender of policyholder (*gender*) and type of fuel (*fuel*). The general statistics are recorded in the next table and outcome distribution is depicted in the figure 1.

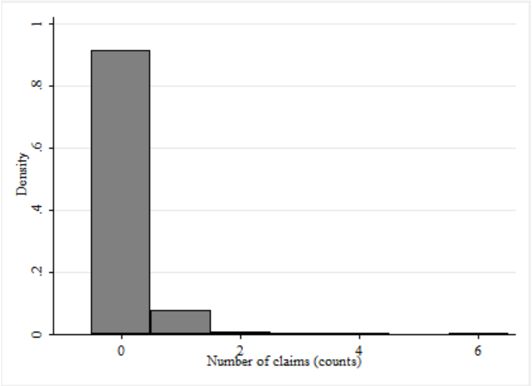
**Tab. 1: General statistics of risk factors considered in the model**

| Variable | Type | Mean | SD | Min | Max | Skew | Kurtosis |
|---|---|---|---|---|---|---|---|
| *agecar* | continuous | 4.4 | 3.4 | 0 | 40 | 0.7 | 3.7 |
| *volume* | continuous | 1 693.2 | 673.7 | 0 | 15 825 | 5.2 | 66.3 |
| *kw* | continuous | 76.8 | 41.1 | 0 | 2 455 | 13.5 | 680.3 |
| *ageman* | continuous | 32.1 | 25.5 | 0 | 90 | -0.1 | 1.6 |
| *nocit* | continuous | 503 951 | 510 794 | 41 255 | 1 249 026 | 0.7 | 1.6 |
| *company* | binary | .6445 | .4787 | 0 | 1 | - | - |
| *gender* | binary | .2118 | .4086 | 0 | 1 | - | - |
| *fuel* | categorical | .3290 | .5631 | 0 | 2 | - | - |

Source: STATA 12.1

**Fig. 1: Outcome distribution (number of claims)**

| Number of claims (counts) | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | 56,654 | 91.53 | 91.53 |
| 1 | 4,637 | 7.49 | 99.02 |
| 2 | 546 | 0.88 | 99.9 |
| 3 | 56 | 0.09 | 99.99 |
| 4 | 5 | 0.01 | 100 |
| 6 | 2 | 0 | 100 |
| **Total** | **61,900** | **100** | |



Source: STATA 12.1

## 2.1 Parameter estimates of selected count models

First and foremost, the Poisson model was estimated as a benchmark model which other models are compared with. During the model fit evaluation, the dispersion statistics (1.117) indicated that data are overdispersed. Therefore, the NB2 model was estimated and it proved that NB2 model is high preferable according to the AIC and BIC criterion. Moreover, the dispersion statistics of NB2 indicates that the overdispersion was eliminated (value is close to 1). However, we may suppose that there are some non-linear relation between some risk factors and the outcome. For that reason and for the suspection of misspecified link function (proved later), we estimated the multivariate fractional polynomial NB2 model to incorporate the non-linearity in the model. We considered the FP of maximum second degree and on the basis of closed test procedure we revealed the most appropriate FP2 tested against the linear and FP1. Then the NB2 model was reestimated. The results in the form of p-values are recorded in the next table.

**Tab. 2: Results of closed test procedure**

| **Type** | **Variable** | **P-value for testing** | | | **Variables and powers selected** |
|---|---|---|---|---|---|
| | | **Inclusion** | **FP2 vs linear** | **FP2 vs FP1** | |
| Continuous | `agecar` | <0.001 | 0.001 | 0.001 | 2; 3 |
| | `volume` | <0.001 | <0.001 | <0.001 | 0.5; 1 |
| | `kw` | <0.001 | <0.001 | <0.001 | 0.5; 1 |
| | `ageman` | <0.001 | <0.001 | 0.121 | -1; 3 |
| | `nocit` | <0.001 | 0.256 | 0.810 | 1 |
| Binary | `gender` | <0.001 | | | |
| | `company` | <0.001 | | | |
| Categorical | `fuel` | <0.001 | | | |

Source: STATA 12.1

According to the results in the table above, we can conclude that the impact of `agecar`, `volume`, `kw` and `ageman` on the outcome is non-linear and modelling the relation via FPs is highly recommended on the 95% confidence interval. In the last column, there are also recorded the optimal powers. We can mention here that `nocit` is modelled via linear function (power 1). The comparison with other models is recorded in the table 3.

**Tab. 3: Comparison of estimated count models**

| **Model** | **AIC** | **BIC** | **alfa** | **dispersion** | **Pseudo $R^2$** |
|---|---|---|---|---|---|
| **Poisson** | 38,762.05 | 38,852.18 | 0.000 | 1.117 | 0.0275 |
| **NB2** | 38,405.52 | 38,495.65 | 1.273 | 0.997 | 0.0247 |

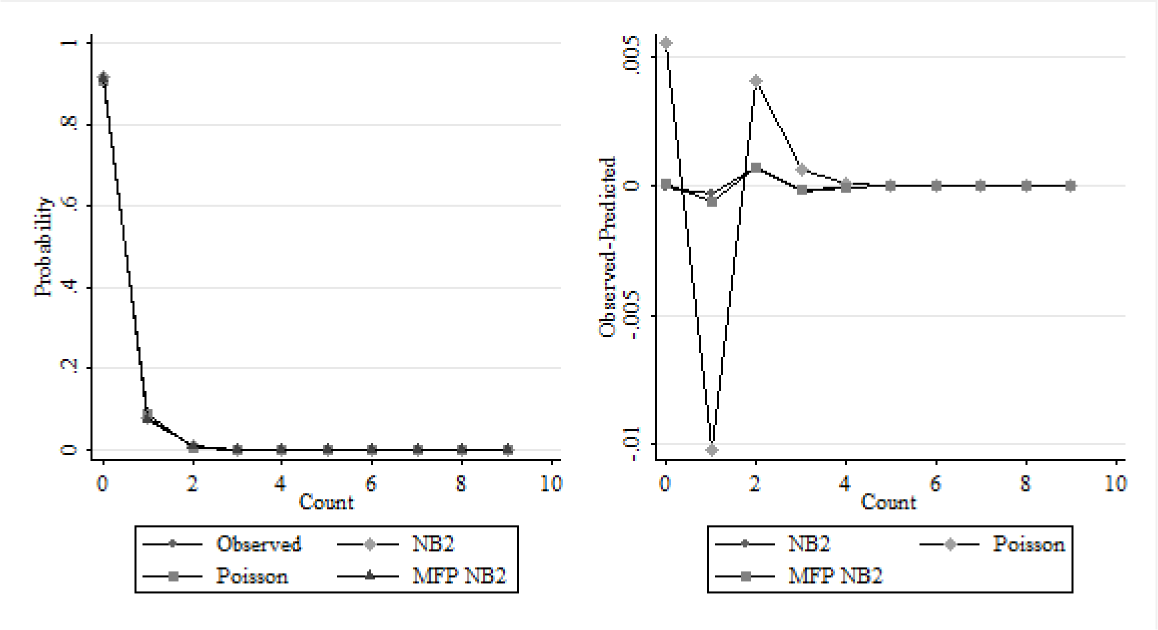
| | | | | | |
|---|---|---|---|---|---|
| **MFP NB2** | 38,298.51 | 38,406.66 | 1.193 | 1.009 | 0.0275 |

Source: STATA 12.1

It is obvious that multivariable FP NB2 model is preferable on the basis of AIC and BIC criterion. The dispersion statistics indicates that the model is still equidispersed and moreover the pseudo coefficient of determination is the same as in the case of Poisson model and higher than when the NB2 model was considered.

Finally, we compare all models and assess the model accuracy. In the next figure, the observed and modelled count probabilities according to each model are depicted.

**Fig. 2: Model accuracy: observed and modelled probabilities (left); difference between observed and modelled probabilities (right)**



Source: STATA 12.1

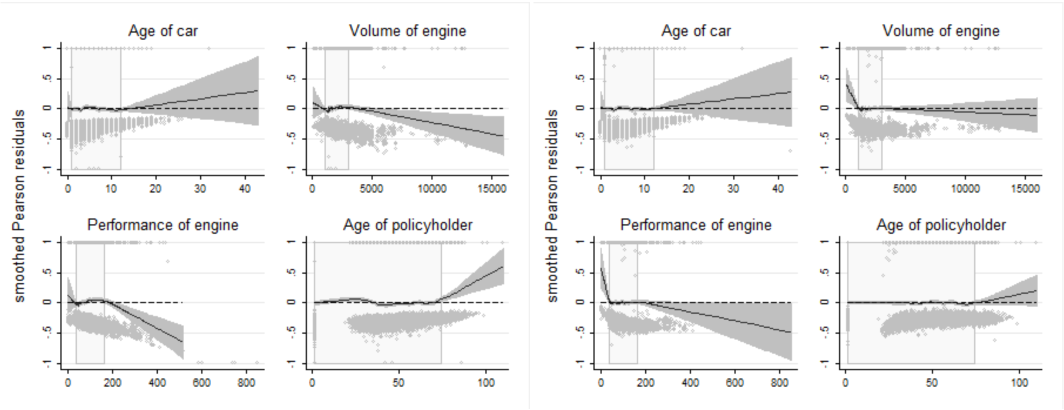
The differences in accuracy of all models are not apparent, therefore we also give the differences between observed and modelled count probabilites. We can see that the Poisson model underestimates the probability of zero and two insured accident significantly and overestimates the probability of one accident. The difference between NB2 and non-linear NB2 model is not noticeable and therefore it must be decided between them on the basis of further analysis.

## 2.2 Residual analysis and assessing of the model fit

To evaluate how the models fit the data, we analyze the Pearson residuals in order to reveal observations well not fitted and in similar manner we analyze hat diagonals. Furthermore, the specification of link function is verified.

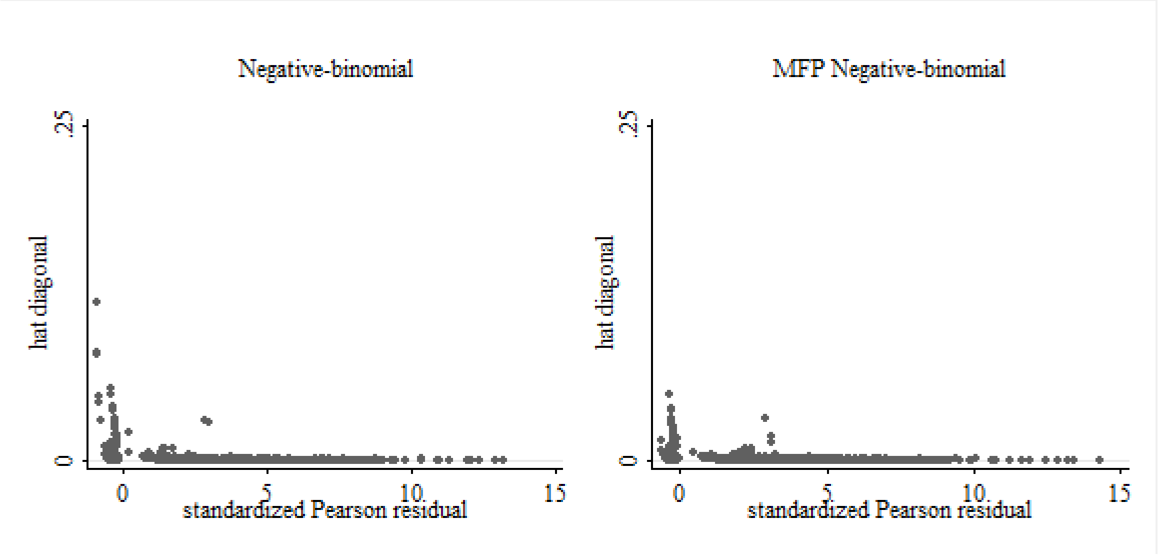
In the next figure, the smoothed residuals of Poisson and NB2 model are depicted (shown only for continuous variables except for `nocit`). We can see that the Poisson model does not fit the data well because the smoothed residuals departure from zero highly, especially at the extremal values (outside the 95% interval of all observations depicted by the gray rectangular). It is apparent that the lack of fit of these observations is reduced in the NB2 model. The difference based on residuals between NB2 and nonlinear NB2 model is negligible (not shown) and further analysis using hat diagonals is necessary, see figure 4 where the hat diagonals are drawn against the standardized Pearson residuals.

**Fig. 3: Smoothed Pearson residuals: (1) Poisson model; (2) NB2 model**



Source: STATA 12.1

**Fig. 4: Residual analysis: (1) NB2 and (2) non-linear NB2 model**



Source: STATA 12.1

All the observations with high hat values indicate the lack of fit. It is apparent that the model fit is improved significantly when the non-linear NB2 model is used (significantly smaller hat values compared to NB2 model). After all, we verified whether the link function had been well specified, see the results in the table 4.

**Tab. 4: Results of link test**

| Statistics | | Poisson | NB2 | MFP NB2 |
|---|---|---|---|---|
| **Linktest p-values** | **hat** | 0.076 | 0.242 | <0.001 |
| | **hat$^2$** | <0.001 | <0.001 | 0.881 |

Source: STATA 12.1

Both Poisson and NB2 model have misspecified link function according to the statistical insignificance of hat and significant squared hat diagonals. As it was mentioned above, neglecting of non-linear relations between risk factors and the outcome may result in misspecified link function which is our case here because the link test indicates that link function is well specified only in non-linear NB2 model.

## Conclusion

The paper was devoted to the proposing of preliminary count model which would be appropriate to modelling of counts and count probabilities of insured accident for given motor-hull insurance portfolio. There were estimated and compared a Poisson model, negative-binomial and nonlinear negative-binomial model. The models comparison was conducted by several statistical tests and by residual analysis.

We proved that the Poisson model is not appropriate for modelling of insured accident probability for given insurance portfolio due to the overdispersion and that the negative-binomial model is preferred. Moreover, we showed that the Poisson model is not really accurate because it overestimates and underestimates the count probabilities. The negative-binomial model suffered by misspecified link function. We handled with this imperfection by using fractional polynomials because we supposed the non-linear relation between some risk factors and the outcome. These nonlinear relations were also proved.

## Acknowledgment

This paper was solved within the project P403/12/P692 Managing and modelling of insurance risks within Solvency II.

## References

Bliss, C. I., & Owen, A. R. G. (1958). Negative binomial distributions with a common k. *Biometricka*, *45*, 37–58.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, *33*(1), 38–44.

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions*. College Station: Stata Press.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge: Cambridge University Press.

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, *15*, 209–225.

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed test procedures with special reference to ordered analysis of variance. *Biometrika*, *76*, 655–660.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, *135*, 370–384.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics*, *29*, 15–24.

Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Applied Statistics*, *43*, 429–467.

Royston, P., Ambler, G., & Sauerbrei, W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, *28*, 964–974.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, *5,* 232–242.

## Contact

Jiri Valecky
VSB-TUO, Faculty of Economics, Department of Finance
Sokolska tr. 33, 701 21 Ostrava
jiri.valecky@vsb.cz