DOI: 10.20472/ES.2020.9.1.012

NOTE ON MISMODELLING OF POLICYHOLDER'S AGE IN CLAIM FREQUENCY MODEL: A MATTER OF GENDER IN VEHICLE INSURANCE

JIŘÍ VALECKÝ

Abstract:

Using the motor hull insurance data of Czech insurer, the paper deals with how mismodelling of policyholder's age can induce misleading conclusions about the gender differences in claim frequency within vehicle insurance. This study is based on individual data with unit policy duration and puts the emphasis on correct modelling of functional form of the age to show that mismodelling as well as categorization yields misleading conclusions and, finally, we demonstrate how the inferences depend on categorization itself. Thus, we showed that linear form as well as the categorization increases the type I error to detect the obvious interaction between gender and age. By involving fractional polynomials, the results partially support the judgement of European Court of Justice to ban using gender as a rating factor, in particular for young policyholders. We concluded that, if another relevant data are not available, the gender as well as interaction with the age should be considered in the claim frequency model although such model cannot be used for setting premium.

Keywords:

age, claim frequency, effect modifier, fractional polynomials, gender, interaction, vehicle insurance

JEL Classification: C31, C51, G22

Authors:

JIŘÍ VALECKÝ, VŠB-TUO, Faculty of Economics, Czech Republic, Email: jiri.valecky@vsb.cz

Citation:

JIŘÍ VALECKÝ (2020). Note on mismodelling of policyholder's age in claim frequency model: a matter of gender in vehicle insurance. International Journal of Economic Sciences, Vol. IX(1), pp. 224-240., 10.20472/ES.2020.9.1.012

1 Introduction

The paper deals with how mismodelling of policyholder's age can induce misleading conclusions about the gender differences in claim frequency within vehicle insurance. In fact, these differences have been observed across the policyholder's age and several studies dealt with this phenomena. However, their conclusions were mostly drawn on average or expected frequencies and involved the grouped data. By contrast, this study is based on individual data with unit policy duration and puts the emphasis on correct modelling of functional form of the age to show how mismodelling suppresses the obvious interaction and how the inferences depend on the categorization of continuous variables. In addition, the study also verifies as well as validates the "gender" effect across the policyholder's age with given level of confidence.

The preference of dataset in which each policy lasts 1 year is motivated by potential bias when grouped data are collected over different time exposure. Suppose two policies for which we observed a claim. One terminates at the end of the year, while the other ended in the middle of the year. Although the claim count is the same, the annual claim frequency of the latter policy is twice higher than the former because the policy was terminated untimely.

Both policyholder's age and gender are relevant risk factor from the statistical point of view. However, the gender might be a proxy for other driver's characteristics. For instance, it is accepted that male drivers pose higher risk than female drivers because of increased aggression, lower limits tolerance and other psychological reasons, see for instance (Cestac et al., 2011) or recently (Harbeck and Glendon, 2018). Further, the different claim frequency is also observed within various subgroups of the policyholders' age as well as differences between males and females across various age. It represents effect modification that requires involving an interaction between gender and age into the claim frequency model.

In these perspectives, the prohibition to use gender for setting premium appears against all empirical observations and statistical results. However, drawing conclusions by comparing average or expected claim frequencies is insufficient and must be supported and confirmed by proper statistical verification. Our results partially support the judgement of European Court of Justice that banned to use the gender as a rating factor since 21.12.2012.

To model the claim frequency, we adopted the concept of generalized linear models (GLMs) where the relationship between the frequency and individual characteristics is expressed via a link function transforming the claim frequency into a linear combination of given individual risk factors. However, the linearity assumption may yield incorrect assessments of effect of policyholder's age on the claim frequency. In addition, as we show in this paper, the obvious interaction with gender is rejected on the 95% confidence level.

Therefore, we involved fractional polynomials (FPs) to specify the functional form correctly. On one hand, these functions are more sensitive to influential observation than linear form because such observation drives not only the value of estimated parameters, but also the degree and power of FP. On the other hand, this approach represents very flexible approach that increases the fit of the data as well as helps to avoid mismodelling when the true function is non-linear.

Further, to avoid estimation of FPs and still to respect a potential non-linearity, continuous variables are sometimes categorized into several groups that allows using the linear model. Then, the non-linearity is handled by non-proportional effect of each successive category.

However, the categorization might be controversial. First, there is a problem how to determine the number of cutpoints and where to place them. The best choice is to use recognized cutpoints that are mostly unavailable. More naturally, the cutpoints are given by percentiles (e.g. quartiles) but it worsens study comparison where each of them is based on different dataset and different percentiles. Second, categorization concerns a loss of information and examination of interaction between categorical and categorized continuous variable yields difficult interpretation of the model due to including many interaction terms. In addition, as we show in this paper, the conclusions about "gender" effect depends on categorization itself that increases the type I error to detect the obvious interaction between gender and age.

Thus, using the motor hull insurance data of Czech insurer, we subject the interaction between policyholder's age and gender to further analysis and we verify it statistically. We show that mismodelling the functional form of policyholder's age as well as its categorization yields misleading conclusions and, finally, we demonstrate how these inferences depend on categorization itself. The remainder of this paper is organized as follows. Section 2 summarizes the literature in this field of study. A claim frequency model based on negative binomial (NB) distribution is described in Section 3, as well as fractional polynomials and effect modification. Section 4 shows how the mismodelling and categorization affects the conclusion about the "gender" effect on claim frequency and presents the conclusions drawn on correctly specified form of policyholder's age. Section 5 gives the conclusions of this study.

2 Literature review

Although the Poisson as well as negative binomial distribution had been known for long time, only development of GLMs by (Nelder and Wedderburn, 1972) put the emphasis on distributional properties and non-linear models that incorporate explanatory predictors.

The first application of GLMs was used to model the claim frequency for marine insurance and the claim size for motor insurance in (McCullagh and Nelder, 1983). More applications of GLMs occurred mostly after the 1990, when the insurance market was being deregulated in many countries and the GLMs were used to undertake a tariff analysis, for example (Andrade-Silva, 1989), (Brockman and Wright, 1992) or (Renshaw, 1994). GLMs are also used for premium optimization, for example (Zaks et al., 2006) or (Branda, 2014), and for the estimation of solvency capital requirement that has appeared recently in (Valecký, 2017).

The first natural choice for modelling count data is Poisson regression model, but it is mostly not sufficient because of overdispersion. Therefore, various types of mixed Poisson model are applied. The negative binomial model was derived from Poisson-gamma mixture distribution, which is now commonly symbolized as NB2, (Cameron and Trivedi, 1986). The comparison of NB with Poisson can be found in (David, 2015). Common alternative for overdispersed data is also quasi-Poisson model, see for instance (Ver Hoef and Boveng, 2007) for comparison with NB model. In addition, the negative binomial model appeared in many extension, for instance as a zero-inflated model, for instance (Kim et al., 2016), or as a generalized model, (Greene, 2008). For review about variations of negative binomial model, we refer to (Hilbe, 2011).

To obtain a well fitted model, it is crucial to identify the relevant factors as emphasized (Kafková and Křivánková, 2014), while (Valecký, 2016) summarized the modelling issues necessary for a good claim frequency model and highlighted the non-linear effect of specific risk factors.

To relax the assumption about the linearity, some authors categorize the continuous factor (Kafková and Křivánková, 2014) or (David, 2015), while the others showed that this approach might be generally dangerous, e.g. (Mazumdar and Glassman, 2000), (Royston et al., 2006)) or (Naggara et al., 2011).

Generalised additive models (GAMs) represents one out of the method to handle the non-linearity. However, we prefer fractional polynomials used extensively by (Royston and Altman, 1994), (Royston et al., 1999), (Sauerbrei and Royston, 1999) because of better interpretation and higher transportability.

The EU Gender Directive of December 13th 2004 (Council Directive 2004/113/EC) provided for equal treatment between men and women in the access and supply of goods and services. However, an exception in Article 5(2) allowed the proportional difference in insurance premiums. Then the judgement of EU's Court of Justice prohibited to use the gender as a rating factor since 21.12.2012 although several studies pointed out potential increase in danger of adverse selection as well as morale hazard, e.g. (Oxera, 2010, 2011). For all that, the gender is still involved as a relevant risk factor in models for the purpose of modelling claim frequency in vehicle insurance, e.g. (David, 2015), (Hsu et al., 2016) or (Summun et al., 2018), even though (Ayuso et al., 2016) or (Verbelen et al., 2018) proved that gender is a proxy for another driver's characteristics, such as experience or driving habits.

3 Negative binomial model

Because the Poisson model does not accommodate the overdispersion, we used the negative binomial model. Let's suppose negative binomial distribution with probability mass function in the form of

$$f\left(y_i; w_i, \kappa, \lambda\right) = \frac{\Gamma\left(w/\kappa + w_i y_i\right)}{\Gamma\left(w/\kappa\right)\Gamma\left(w_i y_i + 1\right)}\left(\frac{1}{\kappa\mu + 1}\right)^{w/\kappa}\left(\frac{\kappa\mu}{\kappa\mu + 1}\right)^{w_i y_i} \tag{1}$$

or in exponential form

$$\begin{aligned} f\left(y_i\right) \quad &= \exp\left\{ w_i\left(y_i \log\left(\frac{\kappa\mu_i}{\kappa\mu_i + 1}\right) - \frac{1}{\kappa}\log\left(\kappa\mu_i + 1\right)\right) + \ldots \right. \\ &\left. \ldots + \log\Gamma\left(\frac{w_i}{\kappa} + w_i y_i\right) - \log\Gamma\left(\frac{w_i}{\kappa}\right) - \log\Gamma\left(w_i y_i + 1\right)\right\}, \end{aligned} \tag{2}$$

where κ is the negative binomial heterogeneity or overdispersion parameter, w_i is the exposure, y_i is the observed claim frequency and μ_i is the mean response. In terms of exponential dispersion model, $\frac{1}{\kappa}\log\left(\kappa\mu_i + 1\right)$ represents the cumulant $b(\theta)$ that is assumed to be twice continuously differentiable with an invertible first derivative because these determine the mean and the variance function as follows

$$E[y_i] = b'(\theta) = \mu_i \,, \tag{3}$$

$$Var(y_i) = \phi \frac{b''(\theta)}{w_i} = \frac{\mu_i + \kappa\mu_i^2}{w_i} \,. \tag{4}$$

Letting $\mu_i = \exp(\eta_i)$, it yields the non-canonical negative binomial model that is referred as NB2 and where η_i represents the systematic component

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} = \mathbf{x}_i\boldsymbol{\beta} \,, \tag{5}$$

where x_{ij} is an observed value on the variable x_j for each policy $i = 1,\ldots,n$ and β are unknown parameters to be estimated.

To obtain the estimates of β, the maximum likelihood Newton-Raphson type algorithm is preferred to IRLS method because observed and expected information matrices are not equivalent in NB2 model and it is generally assumed for non-canonical models that observed standard errors are less biased than expected standard errors.

Further, the log-likelihood function is obtained by substituting $\exp(\mathbf{x}_i\boldsymbol{\beta})$ for each instance of μ in (2), thus

$$\begin{aligned} \ell(\boldsymbol{\beta}, \kappa; y) = \sum_{i=1}^{n} w_i \left\{ y_i \log\left(\frac{\kappa \exp(\mathbf{x}_i\boldsymbol{\beta})}{\kappa \exp(\mathbf{x}_i\boldsymbol{\beta}) + 1} \right) - \left(\frac{1}{\kappa} \log\left(\kappa \exp(\mathbf{x}_i\boldsymbol{\beta}) + 1 \right) \right) \right\} + \ldots \\ \ldots + \log\Gamma\left(\frac{w_i}{\kappa} + w_i y_i \right) - \log\Gamma\left(\frac{w_i}{\kappa} \right) - \log\Gamma(w_i y_i + 1) \Big\}. \end{aligned} \tag{6}$$

The goodness of fit test is performed using the deviance statistic in the form of

$$D = 2\left(\ell(y) - \ell(\mu)\right), \tag{7}$$

where $\ell(y_i)$ and $\ell(\mu_i)$ is the log-likelihood of the saturated model and the fitted model respectively. The value of D is approximately chi-squared distributed with $(n - J - 1)$ degrees of freedom, where n and J represents the number of policies and factors.

The deviance statistic is also used to perform the likelihood ratio (LR) test for model comparison. The statistic is calculated as

$$\Delta D = -2\ell_1\left(\mu^{(1)}\right) + \left\{ -2\ell_0\left(\mu^{(0)}\right) \right\}, \tag{8}$$

where $-2\ell_1\left(\mu^{(1)}\right)$ and $-2\ell_0\left(\mu^{(0)}\right)$ are the log-likelihood of nested and full model. The test statistic is approximately chi-squared distributed with $J_0 - J_1$ degrees of freedom if the models have J_0 and J_1 parameters and if the condition $J_0 > J_1$ holds.

3.1Fractional polynomials

The effect of risk factor on the systematic component is not necessary linear and some transformation is required. One out of the techniques used to handle the non-linearity involves fractional polynomials. Let the expression (5) be rewritten in the form of

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{J}\sum_{k=1}^{m_j}\beta_{jk}F_{jk}(x_{ij}), \tag{9}$$

where $F_{jk}(x_{ij})$ is a particular type of power function and m_j is the degree of FP with powers $p_1, p_2, \ldots, p_m$, denoted as FPm$(p_1\ p_2 \ldots p_m)$. The powers could be any number, but estimation of general powers requires non-linear optimization, which may cause problem with convergence. Therefore, (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

$$F_{jk}(x_{ij}) = \begin{cases} x_{ij}^{p_k}, & p_k \neq p_{k-1}, \\ F_{k-1}(x_{ij})\ln(x_{ij}), & p_k = p_{k-1}, \end{cases} \tag{10}$$

for $k = 1, \ldots, m_j$ and restricting powers p_k to those in S.

Thus, for given degree of FP, all models are estimated and the best model is represented by the highest value of log-likelihood function. Note that, for one variable, the set S generates 8 models for FP1, 36 models for FP2, and so on, yielding a great flexibility. However, the functions of degree $m > 2$ are rarely used.

Next, the variables entering into the model can influence each other as well as the degree and powers of FPs, therefore (Sauerbrei and Royston, 1999) originated a routine for FP selection in multivariable framework, that is called as a MFP algorithm with these steps, see the reference for the original version,

1. The full linear model is fitted.
2. Let $c = 0$, to initialize the cycle counter.
3. Let $j = 1$ to initialize the variables counter within each cycle.
4. All other variables are currently included in the model as adjustment terms. If the x_j is a continuous variable, run the subroutine to identify best FP based on closed-test procedure. Otherwise, test at the α_1 level if to include x_j in the model and skip the following closed-test subroutine.
   i) Test the model with the best FP2 for x_j against the full model with linear function for x_j at the α_2 level using three d.f. If the test is not significant, use the linear function and stop this subroutine, otherwise continue.

ii) Test the model with the best FP2 for x_j against the model with the best FP1 at the α_2 level using two d.f. If the test is not significant, use FP1 and stop the subroutine. Otherwise use the best FP2 for x_j.

5. Let $j = j+1$. If j is smaller or equal to the number of variables, process the next factor (step 4).
6. Let $c = c+1$ and repeat the whole procedure until the convergence.

3.2 Evaluation of effect modifiers

Some risk factors in the systematic component (5) may interact, i.e. one modifies the relationship between the outcome and the other explaining variable entered into the model.

Let us consider the interaction between binary variable x_1 (gender) and continuous variable x_2 (policyholder's age). Their interaction can be expressed as the changing slope parameter of x_1 that depends on the level of x_2, thus

$$\eta = \beta_1 x_1 + \beta_2 x_1 x_2 = (\beta_1 + \beta_2 x_2) x_1 = \beta_{x_1 \times x_2} x_1 \text{ for all } x_2. \tag{11}$$

Clearly, 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 effect of x_1 is ambiguous. In addition, the statistical significance cannot be tested using the z-test (considering the coefficient β_2) because the standard error of the estimated changing parameter $\beta_{x_1 \times x_2}$ is in fact computed as

$$SE_{x_1 \times x_2} = \sqrt{x_1^2 \text{var}(\beta_1) + (x_1 x_2)^2 \text{var}(\beta_2) + 2x_1^2 x_2 \text{cov}(\beta_1; \beta_2)}. \tag{12}$$

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

Thus, the model including the interaction is usually defined as follows

$$\eta = \beta_0 + \beta_1 x_1 + \beta_{02} x_2 I(x_1 = 0) + \beta_{12} x_2 I(x_1 = 1), \tag{13}$$

where $I(\text{condition}) = 1$ if condition holds, 0 otherwise. The model is tested against the nested model

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{14}$$

by LR test, which statistic follows chi-squared distribution with one degree of freedom.

Then, the effect of x_2 (policyholder's age) conditioned on x_1 (gender) is given by $\eta_0(x_2) = \beta_{02} x_2$ and $\eta_1(x_2) = \beta_1 + \beta_{12} x_2$ for males and females respectively, yielding the "gender" effect

$$g(x_1) = \eta_1(x_2) - \eta_0(x_2) = \beta_1 + (\beta_{12} - \beta_{02}) x_2. \tag{15}$$

4 Assessment of "gender" effect across the policyholder's age

In this section, using the motor hull insurance data of Czech insurer, we demonstrated how the conclusions might be manipulated using the methodology and how important is to model the functional form correctly. We firstly presented a model with linear systematic component and a model with categorized policyholder's age, in which we showed how the inferences depend on categorization itself. At the end, we evaluated the effect of gender across the age properly by model that involves FPs. We also verified statistically that the age modifies the effect of gender and vice versa and we additionally performed a partial internal validation to support the importance of the conclusions.

We used the individual data that encompassed the characteristics of policies during the years 2004-2010 (74,721 observations) and these following risk factors were considered: age of car (`agecar`); engine displacement in cm3 (`volume`) and engine power in kW (`kW`); policyholder's age (`ageman`); car value (`value`); number of citizens in a region (`nocit`); gender of policyholder (`gender`; 0 – male, 1 - female); district area (`district`; 14 various regions in Czech republic); and type of fuel (`fuel`; 0 – petrol, 1 - diesel). Remind that each policy had the unit duration.

Finally, note that engine displacement is also one of the key factor of the engine power. Therefore, both `volume` and `kw` cannot be used together in the model because of high mutual dependence indicated by the correlation coefficient of 0.8348. Thus, we defined a new variable (`kwvol`) that combines `volume` and `kw` as a ratio of engine power in kW and engine displacement as follows: $\texttt{kw}/\texttt{volume}\cdot 1000$.

4.1 Model with linear systematic component

First, we show how conclusions might be influenced when the policyholder's age is mismodelled. We estimated the model with linear systematic component and, thereafter, we extended the model by adding the interaction `gender` × `ageman` to assess the effect of gender across the age.

Let $x_1 = \texttt{gender}$, $x_2 = \texttt{ageman}$ and let x^* be other adjustment variables, the estimated model is

$$\eta\left(\mathbf{x}^*, x_1, x_2\right) = -2.7338 - 0.0109 x_{02} - 0.0085 x_{12} + 0.1104 x_1 + \mathbf{x}^* \beta^*,$$

where

$$x_{02} = x_2 I\left(\texttt{gender} = \text{male}\right), x_{12} = x_2 I\left(\texttt{gender} = \text{female}\right),$$

and $I\left(\text{condition}\right) = 1$ if condition is true, 0 otherwise.

The coefficient -0.0109 and -0.0085 represents the conditional effect of `ageman` for males and females respectively, indicating that the systematic component as well as claim frequency is decreasing as age increases. In addition, the coefficient on x_{02}, which is smaller than on x_{12}, indicates that difference by gender is increasing.

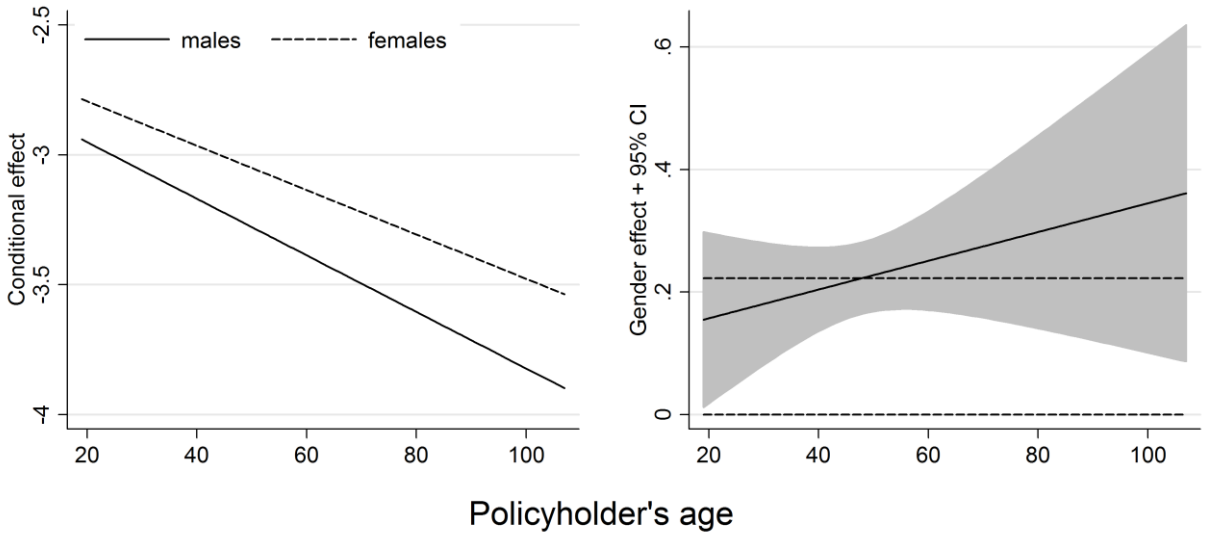
Note that the positive coefficient 0.1104 on `gender` does not indicate itself that the claim is generally more likely from females rather than from males because it also depends on the age

itself. Therefore, we calculated the "gender" effect that represents the varying difference between both gender categories across the age, thus

$$g(x_1) = -0.0085x_2 - (-0.0109x_2) + 0.1104 = 0.0024x_2 + 0.1104,$$

which confirms the increasing difference in claim frequency across gender as the age increases. Next figure shows the estimated conditional effects as well as the "gender" effect with 95% confidence interval.

Figure 1 Model with linear systematic component. Left panel: estimated effects of age in each gender group. Right panel: gender effect function with 95% confidence interval. Horizontal dash line represents zero and the main effect of gender in a model excluding interaction.



Source: Own based on STATA 12

Clearly, the function representing the conditional effect of age for males has higher negative slope that yields the increasing difference in claim frequency from female drivers. It coincides with the increasing "gender" effect shown in the right figure.

Although female drivers are evaluated more risky in general, the "gender" effect suggests that the difference by gender is smaller for young drivers than for middle-aged. However, considering confidence interval of that effect, it does not differ from the main effect of the model without interaction significantly, indicating that the interaction is not statistically significant. It was also confirmed by likelihood ratio test, which yielded a chi-squared value, with one degree of freedom, of 1.03 and the corresponding p-value of 0.3105.

Thus, using the model with linear systematic component, we would conclude that females report claim more often than males regardless of the age and that the difference in claim frequency by gender does not vary significantly across the age. In this case, the variable gender would be considered as confounder rather than effect modifier. However, we showed further that the effect of age is in fact mismodelled, suppressing this effect modification.

4.2 Model with categorized age

In next step, we categorized the policyholder's age to handle its non-linear effect. The categorization is popular because it relaxes the assumption about the linear systematic

component and the model remains easily interpretable. It is used frequently to perform a tariff analysis in which the policies are grouped into tariff cells, associating policies with similar values of risk factors. However, there is no rule for categorization.

First, we grouped the age using the quartiles, i.e. 18-40, 41-51, 52-60, above 60 years, yielding a new variable `catage`. Again, we estimated the model involving this categorized variable and we extended this model by interaction between `gender` and `catage`, keeping all other adjustment variables in the model. Let $x_1 = \texttt{gender}$ and let x_j be the second, third and fourth age category, i.e. for $j = 2,3,4$, we obtained the model

$$\begin{aligned}\eta\left(\mathbf{x}^*, x_1, x_2, x_3, x_4\right) = &-3.0703 - 0.1971x_{02} - 0.1378x_{12} \\ &-0.3334x_{03} - 0.2443x_{13} \\ &-0.3599x_{04} - 0.2307x_{14} + 0.1701x_1 + \mathbf{x}^*\beta^*,\end{aligned}$$

where

$$\begin{aligned}x_{02} &= I\left(\texttt{catage} = 41-51, \texttt{gender} = \text{male},\right), & x_{12} &= I\left(\texttt{catage} = 41-51, \texttt{gender} = \text{female}\right), \\ x_{03} &= I\left(\texttt{catage} = 52-60, \texttt{gender} = \text{male},\right), & x_{13} &= I\left(\texttt{catage} = 52-60, \texttt{gender} = \text{female}\right), \\ x_{04} &= I\left(\texttt{catage} = 61+, \texttt{gender} = \text{male},\right), & x_{14} &= I\left(\texttt{catage} = 61+, \texttt{gender} = \text{female}\right).\end{aligned}$$

Note that the "gender" effect in the first age category is determined by the estimated constant and coefficient on `gender`. Thus, the systematic component as well as claim frequency for males in the first age category is given by the constant -3.0703 and for females of the same category by -3.0703 + 0.1701, indicating that the claim is more likely from females at this age rather than from males. The other estimated coefficients indicate that the claim is less likely as the age category increases regardless the gender (except for the last age group of females). However, the gender difference across the age is varying. Conditional effect of age for both gender is as follows

$$\begin{aligned}&\eta_0\left(x_2\right) = -0.1971x_2 - 0.3334x_3 - 0.3599x_4, && \eta_0\left(x_2\right) = -0.1971x_2 - 0.3334x_3 - 0.3599x_4, \\ &\eta_1\left(x_2\right) = -0.1378x_2 - 0.2443x_3 - 0.2307x_4 + 0.1701, && \eta_1\left(x_2\right) = -0.1378x_2 - 0.2443x_3 - 0.2307x_4 + 0.1701,\end{aligned}$$

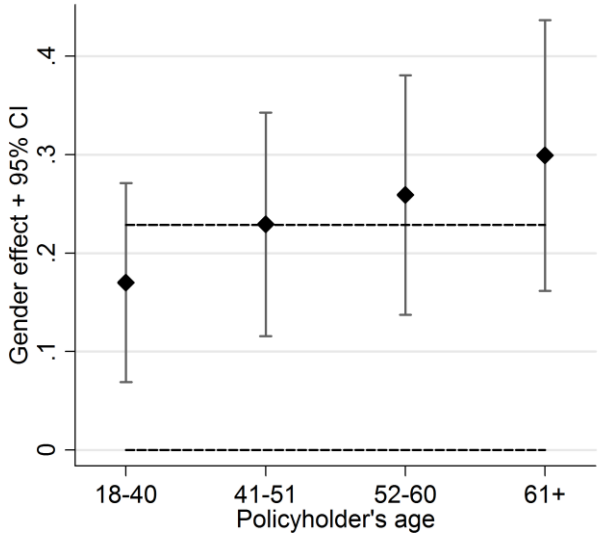
which yields the "gender" effect

$$g\left(x_1\right) = \eta_1\left(x_2\right) - \eta_0\left(x_2\right) = 0.0593x_2 + 0.0891x_3 + 0.1292x_4 + 0.1701.$$

Note that all coefficient are positive, implying that females are more risky for the insurer in each age category. In addition, each coefficient on succeeding age category is higher than the coefficient on preceeding group, indicating the increasing difference by gender.

Unfortunately, although we tried to handle the non-linear effect of age, the estimated effect of gender is not statistical significant and conclusions coincide to these that we drew from the model with linear component. Next figure represents the "gender" effect for all age category including the 95 % confidence interval.

Figure 2 Gender effect across various age group including 95% confidence interval. Horizontal dash line represents zero and the main effect of gender in a model excluding interaction.

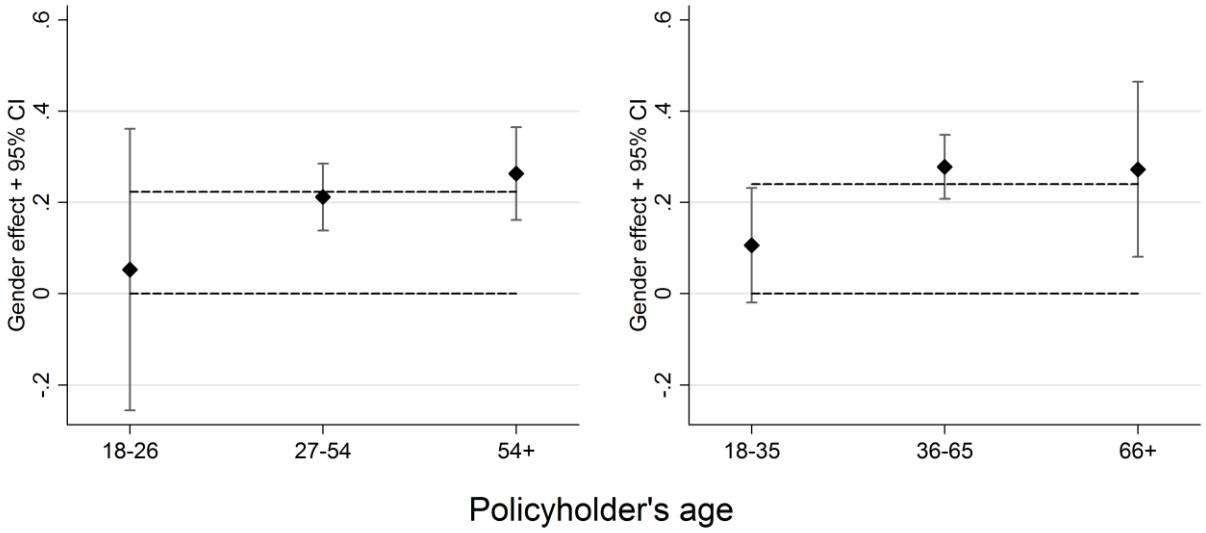


Source: Own based on STATA 12

The figure shows clearly the increasing systematic component for females. However, none of the "gender" effect did not differ significantly from the main effect, indicating that the interaction is not significant either if the non-linearity is handled by categorization. Thus, we would conclude again that the claim frequency does not differ for both gender category across the age. The LR test yielded the chi-squared value of 2.55 (with three degrees of freedom) that corresponds to the p-value 0.1106.

Let's show why the categorization should be avoided. Suppose two another categorization patterns: 1) 18-26, 27-54 and above 54 years; 2) 18-35, 36-65 and above 65. The estimated "gender" effects of both model are shown in next figure.

Figure 3 Gender effect across various age group patterns including 95% confidence interval. Horizontal dash lines represent zero and the main effect of gender in a model excluding interaction.



Source: Own based on STATA 12

Figure shows that using these different categorization patterns yields different conclusions. In contrast to the right figure, the left figure shows clearly that the interaction is insignificant with 95% confidence level, while the second categorization pattern indicates that the claim frequency is lower than frequency given by the main effect with 95% confidence. It was also confirmed by LR test, providing the significance 0.1723 and 0.0172, which corresponds to the chi-squared value (with 2 degrees of freedom) of 1.86 and 5.67 respectively.

In addition, using these two categorization patterns, we found that there is no statistical difference by gender in claim frequency for the first age category. Males and females are statistically equally risky for the insurer, while there was a significant difference by gender when we considered categorization by quartiles. Obviously, the conclusions should not depend on the categorization, which proves that the categorization might yield misleading conclusions.

4.3 Model involving fractional polynomials

Finally, we involved FPs to handle the non-linearity and to avoid the categorization that incurs the loss of information available in the dataset. We applied the MFP algorithm, in which we obtained FP powers of (3 3) for `kwvol`, (-1 3) for `ageman`, (0.5) for `value`, (2) for `agecar`, and a linear term for `nocit`.

Thereafter, the model was extended by adding the interaction `gender × ageman`. Because the MFP applies the centring and scaling on continuous variables (not relevant to the conclusion), let $x_1 = \texttt{gender}$, $x_2 = \frac{\texttt{ageman}}{100}$ and $\bar{x}_2 = mean(x_2)/100$. The other adjustment variables $\mathbf{x}^*$ were also transformed and entered into the model. Thus, we obtained the model

$$\eta(\mathbf{x}^*, x_1, x_2) = -2.6914 + 0.4826x_{01} + 1.5552x_{02} + 0.2699x_{11} + 0.7240x_{12} + 0.2989x_1 + \mathbf{x}^*\beta^*,$$

where

$$x_{01} = \left(x_2^{-1} - \bar{x}_2^{-1}\right) I(\texttt{gender} = \text{male}), \quad x_{11} = \left(x_2^{-1} - \bar{x}_2^{-1}\right) I(\texttt{gender} = \text{female}),$$
$$x_{02} = \left(x_2^{3} - \bar{x}_2^{3}\right) I(\texttt{gender} = \text{male}), \quad x_{12} = \left(x_2^{3} - \bar{x}_2^{3}\right) I(\texttt{gender} = \text{female}),$$

and $I(\text{condition}) = 1$ if condition is true, 0 otherwise.

Next, we tested the model against the nested model without interaction using likelihood ratio test, which yielded a chi-squared value of 50.22 with two degrees of freedom. The corresponding p-value was less than 0.00001, indicating that the interaction between `gender × ageman` is significant.

Further, the effect of age conditioned on gender is given by the model as follows

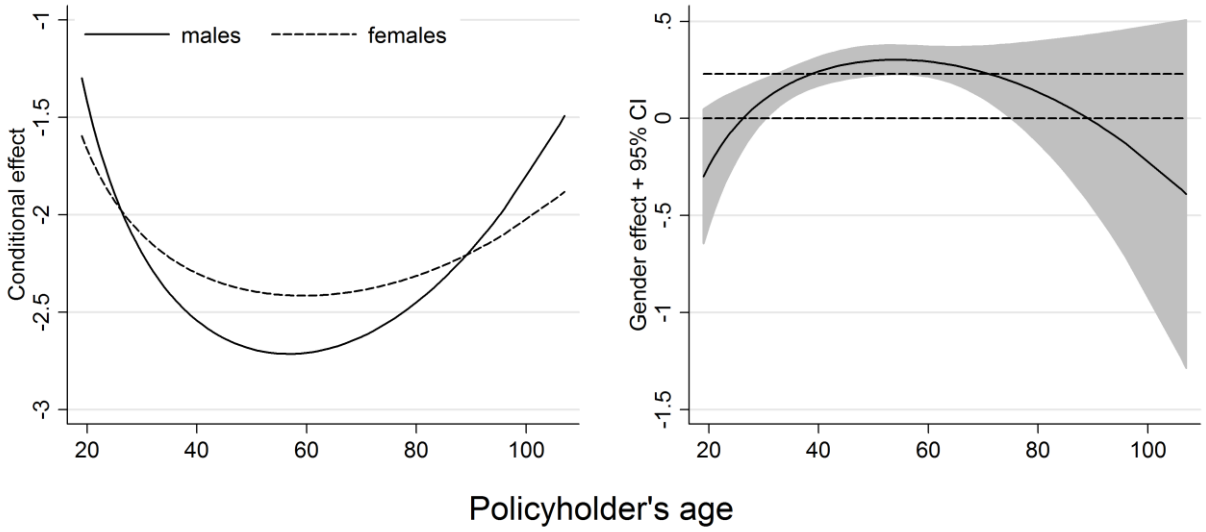
$$\eta_0(x_2) = 0.4826\left(x_2^{-1} - \bar{x}_2^{-1}\right) + 1.5552\left(x_2^3 - \bar{x}_2^3\right),$$
$$\eta_1(x_2) = 0.2699\left(x_2^{-1} - \bar{x}_2^{-1}\right) + 0.7240\left(x_2^3 - \bar{x}_2^3\right) + 0.2989,$$

which yields the "gender" effect

$$g(x_1) = \eta_1(x_1) - \eta_0(x_1) = -0.2127\left(x_2^{-1} - \bar{x}_2^{-1}\right) - 0.8312\left(x_2^3 - \bar{x}_2^3\right) + 0.2989.$$

Because of non-linear transformation of age as well as due to centring and scaling, the interpretation of conditional effects as well as "gender" effect is represented by next figure.

Figure 4 Model involving FPs. Left panel: estimated effects of age in each gender group. Right panel: gender effect function with 95% confidence interval. Horizontal dash line represents zero and the main effect of gender in a model excluding interaction.

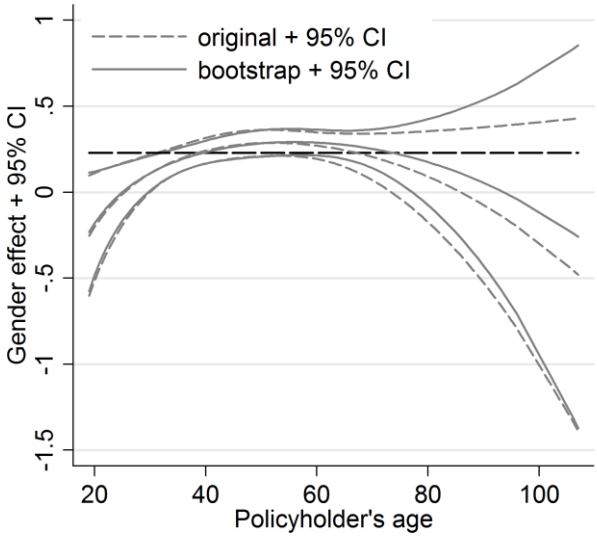


Source: Own based on STATA 12

The left figure shows clearly that claims are less likely from young women than young men in the 18-26 years interval, whereas they are more likely over 26 years. However, the "gender" effect plot indicates that the effect of gender changed from negative to positive and that it significantly differs from the main effect on the interval 18-32 years. More important, it is obvious that there is no statistical difference between males and females within the 18-30 years interval as well as above 75 years, while claims are more likely from women in the interval 31-74 years. Thus, although we observed higher frequency for young men, there is no statistical difference from young women. Significant distinction is confirmed in the middle age only, i.e. above 31 years.

Finally, although the interaction `gender × ageman` was statistically confirmed, the spurious interaction may be of concern because the effect modification might be data-driven. Because we did not have external data, we performed a partial internal validation by bootstrap resampling. We generated 100 bootstrap samples and we estimated the "gender" effect for each. Next figure compares the original and bootstrap "gender" effect functions.

Figure 5 Mean and 95% confidence interval of gender effect function from 100 bootstrap replications and gender effect function on the original data with 95% confidence interval.



Source: Own based on STATA 12

Clearly, even if the data were changed randomly with replacement, significant “gender" effect was detected. The original and the bootstrap “gender” effect as well as both relevant 95% confidence intervals differ only for `ageman` above 60 years, indicating that the “gender” effect function is data-driven on this interval. However, they also confirmed no difference in claim frequency by gender in the 18-30 years interval, while significance difference appeared above 30 years.

5 Conclusion

Using the motor hull insurance data of Czech insurer, the paper demonstrated how the conclusions might be manipulated using the methodology and how important is to model the functional form correctly.

The study showed that mismodelling of policyholder’s age induce misleading conclusions about the gender differences in claim frequency. Assuming the linear form for the age, the gender was not identified as an effect modifier for the age with 95% confidence. It was also confirmed that categorization incurs the loss of information and that the interaction between age and gender was not detected although the non-linear effect of age was treated as non-proportional effect of age groups. It implies that linear form as well as the categorization increase the type I error to detect the obvious interaction between gender and age. In addition, using different categorization patterns, we showed that the significance of effect modification might depend on the categorization and yield contrary conclusions about the effect modification.

Involving fractional polynomials confirmed that gender is significant effect modifier for the age, in particular for young policyholders. In addition, the study also validates the “gender” effect across the policyholder’s age with 95% confidence. In these perspectives, the EU’s ban to use gender for setting premium appeared reasonable at least for young drivers. On the other hand, significant differences in claim frequency by gender appeared and were validated for the age above 30.

However, it does not necessary imply a gender disparity. As others showed, the gender represents a proxy for another characteristic, such as experience, driving habits, etc., and the question is how the insurer will deal with this problem. For instance, one out of the insurance companies in Czech republic started to set the premium for vehicle insurance according to the annual mileage.

Thus, we may conclude that gender should be entered into the claim frequency model at least as a proxy if another relevant data are not available. However, such model cannot be used for setting premium. It also implies that the interaction between gender and age should be considered in such frequency model. In addition, the linear form of the policyholder's age must be carefully verified otherwise some proper technique to handle non-linearity of the age, such as fractional polynomials, should be involved and, finally, categorization should be avoided because the approximation of non-linear effect is insufficient.

Acknowledgements

This paper was supported within Operational Programme Education for Competitiveness – Project No. CZ.1.07/2.3.00/20.0296 and under the SGS project SP2018/154.

References

ANDRADE-SILVA, J.M. (1989). An application of Generalized Linear Models to Portuguese Motor Insurance. In *Proceedings XXI ASTIN Colloquium*, New York. pp. 633. https://doi.org/10.3390/risks4020010

AYUSO, M., GUILLEN, M., and MARIA PEREZ-MARIN, A. (2016). Telematics and Gender Discrimination: Some Usage-Based Evidence on Whether Men's Risk of Accidents Differs from Women's. *RISKS*, 4(2).

BRANDA, M. (2014). Optimization Approaches to Multiplicative Tariff of Rates Estimation in Non-Life Insurance. *Asia-Pacific Journal of Operational Research*, 31(5), pp. 1450032. https://doi.org/10.1142/S0217595914500328

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), pp. 457–543. https://doi.org/10.1017/S0020268100019995

CAMERON, A.C. and TRIVEDI, P.K. (1986). Econometric models based on count data: Comparisons and applications of some estimators. *Journal of Applied Econometrics*, 1(1), pp. 29–53. https://doi.org/10.1002/jae.3950010104

CESTAC, J., PARAN, F., and DELHOMME, P. (2011). Young drivers' sensation seeking, subjective norms, and perceived behavioral control and their roles in predicting speeding intention: How risk-taking motivations evolve with gender and driving experience. *SAFETY SCIENCE*, 49(3), pp. 424–432. https://doi.org/10.1016/j.ssci.2010.10.007

DAVID, M. (2015). Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics and Finance*, 20, pp. 147–156. https://doi.org/10.1016/S2212-5671(15)00059-3

GREENE, W. (2008). Functional forms fo the negative binomial model for count data. *Economics Letters*, 99, pp. 585–590. https://doi.org/10.1016/j.econlet.2007.10.015

HARBECK, E.L. and GLENDON, A.I. (2018). Driver prototypes and behavioral willingness: Young driver risk perception and reported engagement in risky driving. *JOURNAL OF SAFETY RESEARCH*, 66, pp. 195–204. https://doi.org/10.1016/j.jsr.2018.07.009

HILBE, J.M. (2011). *Negative binomial regression*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511973420

HSU, Y.-C., CHOU, P.-L., and SHIU, Y.-M. (2016). An examination of the relationship between vehicle insurance purchase and the frequency of accidents. *Asia Pacific Management Review*, 21(4), pp. 231–238. https://doi.org/10.1016/j.apmrv.2016.08.001

KAFKOVÁ, S. and KŘIVÁNKOVÁ, L. (2014). Generalized Linear Models in Vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62(2), pp. 383–388. https://doi.org/10.11118/actaun201462020383

KIM, D.-H., RAMJAN, L.M., and MAK, K.-K. (2016). Prediction of vehicle crashes by drivers' characteristics and past traffic violations in Korea using a zero-inflated negative binomial model. *TRAFFIC INJURY PREVENTION*, 17(1), pp. 86–90. https://doi.org/10.1080/15389588.2015.1033689

MAZUMDAR, M. and GLASSMAN, J.R. (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in Medicine*, 19(1), pp. 113–132. https://doi.org/10.1002/(SICI)1097-0258(20000115)19:1<113::AID-SIM245>3.0.CO;2-O

MCCULLAGH, P. and NELDER, J.A. (1983). *Generalized linear models*. London: Chapman & Hall. https://doi.org/10.1007/978-1-4899-3244-0

NAGGARA, O. et al. (2011). Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms. *American Journal of Neuroradiology*, 32(3), pp. 437–440. https://doi.org/10.3174/ajnr.A2425

NELDER, J.A. and WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, 135(3), pp. 370–384. https://doi.org/10.2307/2344614

OXERA. (2010). The use of gender in insurance pricing. *ABI Research paper*, 24, pp. 1-91.

OXERA. (2011). *The impact of a ban on the use of gender in insurance*. Report from December 7th 2011.

RENSHAW, A.E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin*, 24(2), pp. 265–285. https://doi.org/10.2143/AST.24.2.2005070

ROYSTON, P. and ALTMAN, D.G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modeling. *Applied Statistics*, 43(3), pp. 429–467. https://doi.org/10.2307/2986270

ROYSTON, P., ALTMAN, D.G., and SAUERBREI, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1), pp. 127–141. https://doi.org/10.1002/sim.2331

ROYSTON, P., AMBLER, G., and SAUERBREI, W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28(5), pp. 964–974. https://doi.org/10.1093/ije/28.5.964

SAUERBREI, W. and ROYSTON, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society Series A*, 162(1), pp. 71–94. https://doi.org/10.1111/1467-985X.00122

SUMMUN, K., KHAN, N.M., JANNOO, Z., SUNECHER, Y., and VEERASAMY, I. (2018). An assessment of the determinants of Mauritian automobile insurance claims using negative binomial and gamma regression models. *JOURNAL OF STATISTICS & MANAGEMENT SYSTEMS*, 21(5), pp. 725–740. https://doi.org/10.1080/09720510.2017.1292686

VALECKÝ, J. (2016). Modelling claim frequency in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 64(2), pp. 683–689. https://doi.org/10.11118/actaun201664020683

VALECKÝ, J. (2017). Calculation of SCR on non-life underwriting risk by using individual risk models. *Prague Economic Papers*, 26(4), pp. 450–466. https://doi.org/10.18267/j.pep.621

VER HOEF, J.M. and BOVENG, P.L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11), pp. 2766–2772. https://doi.org/10.1890/07-0043.1

VERBELEN, R., ANTONIO, K., and CLAESKENS, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS*, 67(5), pp. 1275–1304. https://doi.org/10.1111/rssc.12283

ZAKS, Y., FROSTIG, E., and LEVIKSON, B. (2006). Optimal pricing of a heterogeneous portfolio for a given risk level. *ASTIN Bulletin*, 36(1), pp. 161–185. https://doi.org/10.1017/S0515036100014446

