Fractional polynomials analysis of relation between insured accident and selected risk factors

Jiří Valecký[1]

Abstract. The paper is devoted to the examination of impact of selected risk factors on the insured accidence in motor hull insurance in the Czech Republic. In compliance with these risk factors, it is possible to determine the probability of insured accident for each policyholder and also to differentiate the paid insurance premiums which should corresponds to the undertaken risk. Therefore the impact of risk factors on the probability of insured accidence should be determined the most precisely. In this purpose, there are considered many joint variables, for instance volume and performance of engine, age of car, age of policyholder, and others. Firstly, using general univariable logistic regression, the most statistically significant risk factors are identified and it is determined how they affects occurrence of insured accident. It is also verified if the relation between them is linear. In the case of non-linearity identified, the most appropriate fractional polynomial function is obtained in order to improve fitting data. Finally, the non-linear functions are subjected to the stability analysis and the resulted linear and non-linear relations are interpreted.

Keywords: logit model, fractional polynomial, stability analysis, motor hull insurance, insurance premium, insured accident.

AMS Classification: 62 J 12, 62 P 05
JEL Classification: C31, C58, G22

1 Introduction

Insurance premium rate is determined per monetary unit in accordance with undertaken risk. In other words, the more risky client should pay higher premiums. 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 or according to the size of district where the client lives. Moreover, some insurers respect even client´s age. To differentiate premiums in such way, the precise determining the relations between given particular risk factors and the outcome (insured accident) is crucial.

For the purpose of evaluating the undertaken risk represented by of insured accident probability, one can employ several models based on the GLM family model, see [1]. Nevertheless, the binomial models or count models are applied here, see [2] for instance. Some researchers utilize the advantages of both models and concentrate on the hurdle models representing mostly the combination of logit and some count (Poisson or negative-binomial) model designed firstly by [3].

It is necessary to note here that modelling relation between a risk factors and the outcome may suffer many imperfections resulted from excessive simplification. The simplification which we focus on is assumption of linear logit which may lead to the excessive distortion of relation between risk factor and the outcome. Due to this fact we may encourage to suppose that some risk factor affects the outcome differently for various values and thus the relation between them is necessary to model via non-linear function such as fractional polynomials (FPs), see [4] or [5].

The FPs are quite easy to use and interpret. In spite of it they suffer several imperfections. First and foremost, a selection of FPs is generally sensible to outliers and to data used to fitting the outcome. While the first problem represents the fact that one covariates affects the FPs, see [6], the latter incurs a lack of data fit when the estimated model is transferred to another data sample, see [7]. However, the high leverage does not necessary imply a large effect on fitting the outcome and on selecting FPs. Excluding the outlier observation or covariate pattern is needed to assess if the value may be considered as influential. Thus, the jackknife method, [8], is applied to FPs selection procedure and changes of FPs are examined. On the other hand, to assess the stability of estimated FPs is based on resampling and reestimation of the model (bootstrapping). All variously changing powers and degrees of FPs indicates the function instability. Moreover, allowing this type of uncertainty within FPs selection enlarge the confidence interval of the predicted outcome.

[1] VŠB-TU Ostrava, Faculty of Economics, Department of Finance, Sokolská tř. 33, 701 21, jiri.valecky@vsb.cz.

The aim of this paper is to assess the impact of selected risk factors on the insured accident probability and analyze the non-linear relation between them. The paper is organized as follows. The general logistic regression and its extension by incorporating fractional polynomials are described in Section 2. Section 3 is focused on the empirical examination of selected risk factors and Section 4 concludes the paper.

2 Identifying effect of risk factor on insured accident

Next, for the purpose of quantifying relation of risk factors to the insured accident, we focus on the logistic regression and fractional polynomials.

2.1 Linear logistic regression model

Consider a binary variable Y_i characterizing the occurrence of insured accident for given policyholder, thus

$$Y_i = \begin{cases} 1 & \text{if insured accident occurs,} \\ 0 & \text{otherwise, for } i = 1,...,N, \end{cases} \tag{1}$$

where N is the number of policyholders and each client is characterized by the vector of individual K risk factors $\mathbf{x}_i = (1, x_{1i}, x_{2i}, ..., x_{Ki})$.

The probability of an insured accident for a given policyholder, $P_i = P(Y_i = 1)$, is possible to express, on the basis of its characteristic vector $\mathbf{x}_i$, as a function $F(\boldsymbol{\beta}; \mathbf{x}_i)$ which is monotonically increasing $F'(\boldsymbol{\beta}; \mathbf{x}_i) \geq 0$ and has a domain of definition $(-\infty, +\infty)$ and a range $(0,1)$. Thus, it holds that $F(-\infty) = 0$ and $F(+\infty) = 1$, the probability function can be written in the form of

$$P_i = F(\boldsymbol{\beta}; \mathbf{x}_i), \tag{2}$$

where $\boldsymbol{\beta}$ is vector of parameters $(\beta_0, \beta_1, ..., \beta_K)$.

These properties are satisfied by the cumulative distribution function of the logistic distribution

$$P_i = P(Y_i = 1) = F(\boldsymbol{\beta}; \mathbf{x}_i) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}} = \frac{1}{1 + e^{-(\boldsymbol{\beta}'\mathbf{x}_i)}} \tag{3}$$

which is also a function of the probability that insured accident occurs. The probability that the accident does not occur can be written as

$$1 - P_i = P(Y_i = 0) = 1 - F(\boldsymbol{\beta}; \mathbf{x}_i) = \frac{1}{1 + e^{\boldsymbol{\beta}'\mathbf{x}_i}}. \tag{4}$$

The ratio of probabilities (3) and (4) is referred to as odds and it takes the form of

$$\frac{\pi}{1 - \pi} = \frac{P(Y_i = 1)}{P(Y_i = 0)} = e^{\boldsymbol{\beta}'\mathbf{x}_i} \tag{5}$$

and the logarithm of (5) is termed logit or log-odds, thus

$$\ln\left[\frac{\pi}{1 - \pi}\right] = \boldsymbol{\beta}'\mathbf{x}_i = g(\mathbf{x}_i). \tag{6}$$

2.2 Fractional polynomial function

Equation (6) may not necessary be linear. The non-linearity can be dealt with in that way that the predictor is converted into categorical factors but this procedure incorporates a problem arisen from making a cutpoints. The individuals close to but on opposite sides of the cutpoint are characterized very different rather than very similar, see [9]. Let's define the fractional polynomial function and rewrite the Equation (6) into the form of

$$g(x) = \beta_0 + \sum_{j=1}^{J} \beta_{1j} F_j(x_1) + \beta_2 x_2 + ... + \beta_K x_K, \tag{7}$$

where $F_j(x_1)$ is a particular type of power function. The power p_j could be any number, but [4] 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

$$F_j(x_1) = \begin{cases} x_1^{p_j}, & p_j \neq p_{j-1} \\ F_{j-1}(x_1)\ln(x_1), & p_j = p_{j-1}. \end{cases} \tag{8}$$

The identification and comparison of the most appropriate FPs is made by closed test procedure, [10], which is generally preferred over the sequential procedure, [4].

3 Evaluating effects of selected risk factors

In this section, nonlinear impacts of given risk factors on the insured accident are examined. For this purpose, we used 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 continuous 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*) and average age in the region (*avgagereg*).

First and foremost, using the univariable logistic regression, the statistically significant predictors are identified and their linear effect on the outcome is verified by testing against the fractional polynomials considered at maximum second degree. Afterwards, if the non-linear effects are indicated, the FPs being statistically significant are estimated and the appropriateness of the FPs is assessed by residual analysis. Finally, because of the fact that the FPs are sensitive to the influential observations and to the sample used for estimation, we conduct a stability analysis to examine the stability of selected FPs.

3.1 Examination of non-linearity

Using the closed test procedure, the most appropriate FPs of the first and second degree is estimated and tested against each other and against the linear function. The results are recorded in the next table, where the second column informs about the statistical significance of variable inclusion in the model. The third and fourth columns report about the superiority of FP2 over the linear and FP1. Lastly, there are recorded estimated powers.

Variable	P-value for testing			Powers selected
	Inclusion	FP2 vs linear	FP2 vs FP1	
agecar	<0.001	0.097	0.043	1
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	<0.001	-2, -2
nocit	<0.001	0.617	0.510	1
avgagereg	<0.001	0.006	0.003	-2; -2

Table 1 Univariable analysis

It is obvious on the basis of p-values that all variables may be used for the purpose of insured accident modelling. Focusing on the decision about employing FPs, we can recommend FP2 for *volume*, *kw*, *ageman* and *avgagereg*. On the contrary, modelling *agecar* and *nocit* by FP2 is not statistically better and linear function (power selected = 1) is preferred at 5% level. The shapes of all functions are depicted in next figure to interpret the non-linear relation between the risk factor and the outcome.

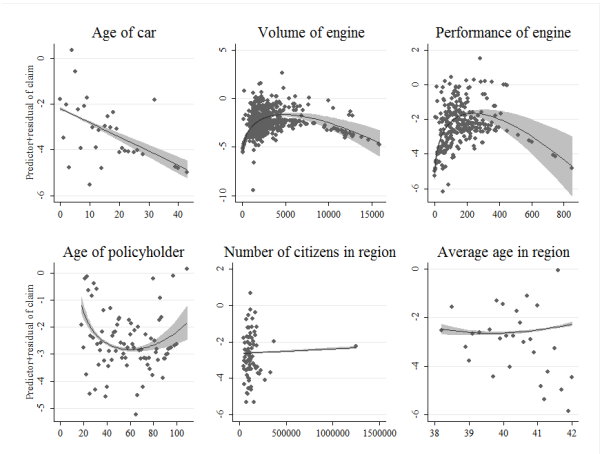

Figure 1 Behavior of functions: residuals and prediction with 95% CI

We can see that the *agecar* and *nocit* are modeled by a linear function, while the polynomials are used to model the other variables. On the basis of the figure above, we can conclude how the selected risk factors influ-

ence the outcome. When the higher the age of car is, the smaller probability of insured accident is. The impact of *volume* and *kw* on the outcome (logit or insured accident probability) is very similar, i.e. the insured accident probability gets higher as both variables increases and descends from some given value. In terms of the impact risk factor on the outcome, effect of *ageman* is also interesting. According to the shape of function, the lower the probability is when the *ageman* is increasing and then gets higher. The behavior of *avgagereg* is similar to the *ageman*. It results from the fact that the drivers around the 40 are the least risky drivers and therefore the region with lower average age indicates the least insured accident probability. Finally, we should note here that *nocit* should be probably modeled as categorical variable or offset rather than continuous variable and only *ageman* or *avgagereg* should be probably used in a multivariable model.

To assess the fact if the FPs improve the model fitting, we compare smoothed Pearson residuals of linear and FP functions for all variables which it is reasonable for, i.e. variables modeled by linear function and recommended as categorical are excluded, see next figure.

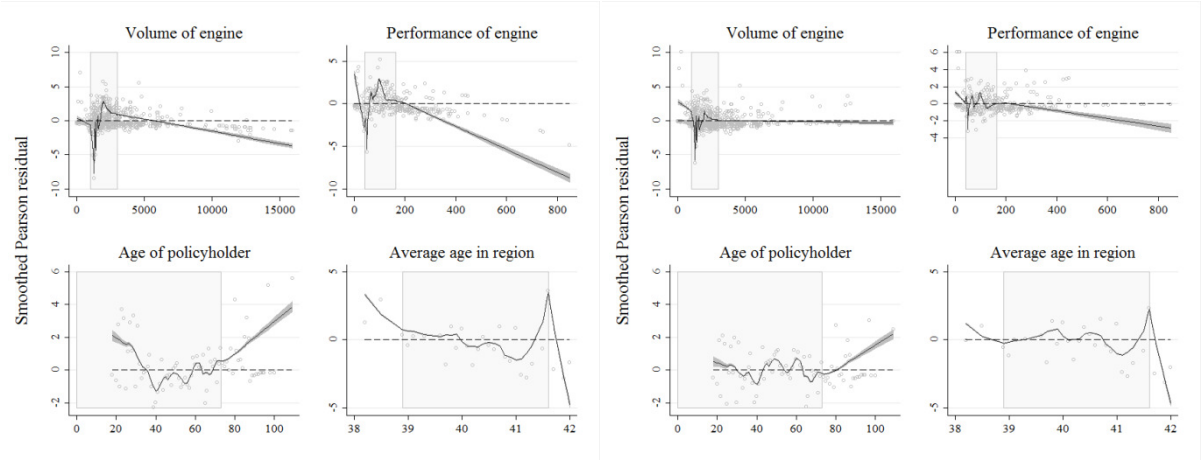

Figure 2 Smoothed Pearson residuals. Left: linear function. Right: FP2 functions

Well fitted observations are indicated by the smoothed residuals close to zero. From the figure above it is apparent that data are fitted better when the relations are modeled via FPs. The lack of fit was reduces in 95 % CI (gray rectangle) of all observations and even outside this region (for outliers).

3.2 Stability analysis of FPs

In the next step, we focus on identifying the influential covariates and in the second step we verify the stability of FPs selected. In the next figure, the calculated deviance difference between linear and FP2 are plotted against the excluded covariate pattern. If the difference is smaller than χ^2 threshold, selection of FP2 depends on this covariates. The plot is amended by the size of circle indicating the covariate frequency to reveal if the variable has several or many observations. The figure 3 gives the evidence that linear function for *ageman* is preferred over the FP2 because excluding all of covariate patterns (except one) results in linear function in FP selection procedure. Using FP2 to model the effect of *volume*, *kw* and *ageman* on the outcome is recommendable because none of the covariates affect the FP selection. Finally, selection of FP2 for modelling *avgagereg* is highly dependent on including the values 41.6 and a little on 38.5.

To assess the stability of selected FPs, we draw one hundred bootstrap sample of size 15 000 and apply the closed-test procedure to determine the most appropriate FPs. For each bootstrap sample is selected the most appropriate FP and the results are depicted in the next figure 4.

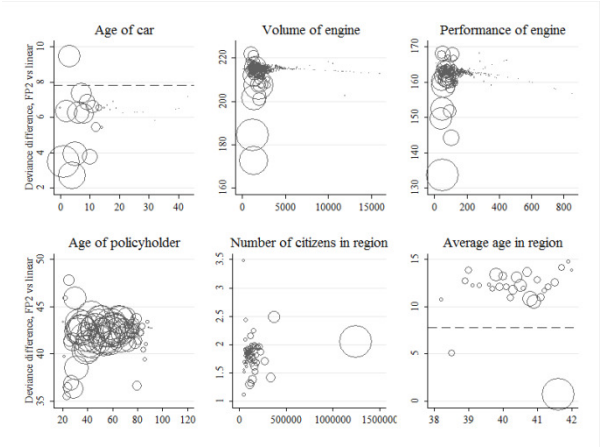

Figure 3 Influential points. Deviance difference between FP2 and linear function for each risk factor excluding each of covariate pattern in turn (circles are proportional to pattern frequencies). Horizontal line at 7.81 represents the χ^2 threshold for significance of FP2 versus linear at the 5% level.

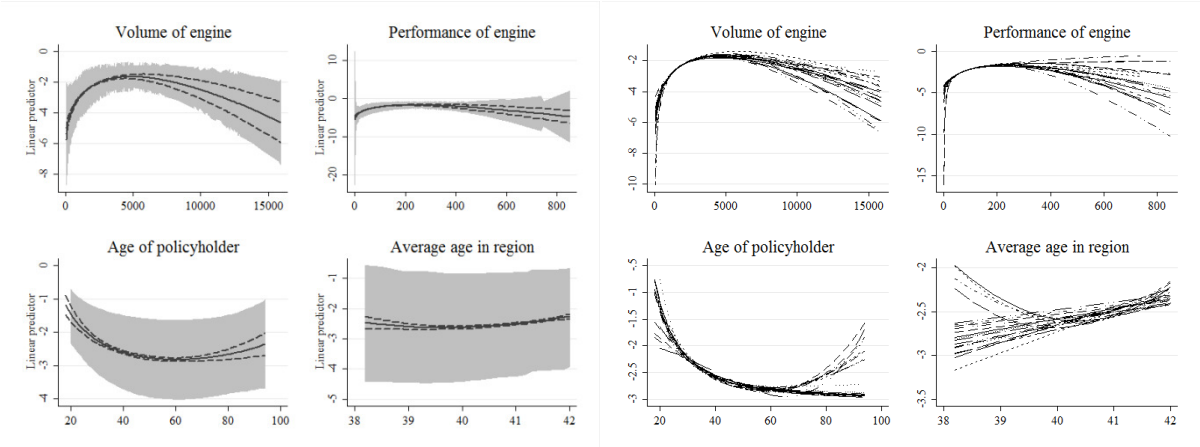

Figure 4 Bootstrap analysis (100 replications) of FPs: Left: Bootstrapped functions with CI: solid lines, means; dashed lines, 95% CI ignoring uncertainty due to FPs selection; gray area, bootstrapped 95% CI. Right: random set of 20 fitted FPs on bootstrap replications.

The predicted outcome can takes any value in 95% CI represented by the dashed line. However, this interval does not incorporate the uncertainty arisen from the selection of FPS using different sample. Therefore the bootstrapped CIs are wider (gray area). We can see that wide CI for *ageman* and *avgagereg* indicates also the high variability of degree and powers of FP. It is more evident in the figure on the right where the random set of 25 FPs out of the 100 bootstrap replications is depicted. We can mention that modelling of *avgagereg* by FP is unstable because of the linear function was selected within some bootstrap samples, probably depending on including the influential covariates, see above. The shape of FP for *ageman* also varies however the descent of probability with increasing age is still apparent. On the contrary, if the probability gets higher from given age of the driver is arguable. The shape of FP for *volume* and *kw* seems to be stable (except for the hook for small *kw*).

4 Conclusion

The paper was devoted to the analyzing impact of selected risk factors on insured accident in given motor hull insurance portfolio. Firstly, the significance of continuous predictors was evaluated within univariable logistic regression and the linear relation was verified. The revealed non-linearity was handled with the fractional polynomials of second degree and the relations were described and interpreted.

On the basis of conducted study, we found all considered continuous predictors to be statistically significant. The relation between age of car and the outcome may be approximated by linear function due to the many influential observations. Moreover, the number of citizens (also average age in region) should be converted into a categorical variable or should be used as an offset in the model rather than a continuous predictor. For others variables the fractional polynomials are recommended. In addition to the results of statistically significance at

the 95% confidence interval, the observations including outliers were fitted better and no other influential observations were identified.

We also revealed that the higher the age of man is, the smaller probability of insured accident is. However, the trend has changed for age above 60 approximately. As the stability analysis confirmed, the latter conclusion is discussable and should be subjected to further analysis because the fractional polynomial selected in the initial univariable analysis was unstable and monotonic decreasing FP was estimated on some bootstrap samples. The impact of volume and performance of the engine on the outcome was very similar, i.e. the insured accident probability gets higher as both variables increases and descends from some given value. The stability of these functions was confirmed except for the hook for low performance.

Acknowledgements

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

References

[1] Altman, D.G., and Andersen, P.K.: Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine* **8** (1989), 771–783.
[2] Cameron, A.C., and Trivedi, P.K.: Econometric models based on count data: Comparisons and applications of some estimators. *Journal of Applied Econometrics* **1** (1986), 29–53.
[3] Marcus, R., Peritz, E., and Gabriel, K.R.: On closed test procedures with special reference to ordered analysis of variance. *Biometrika* **76** (1976), 655–660.
[4] Mosteller, F., and Tukey, J.W.: *Data analysis and regression*. Addison-Wesley, New York, 1977.
[5] Mullahy, J.: Specification and testing of some modified count data models. *Journal of Econometrics* **33** (1986), 341–365.
[6] Nelder, J.A., and Wedderburn, R.W.M.: Generalized linear models. *Journal of the Royal Statistical Society A* **135** (1972), 370–384.
[7] Royston, P., and Altman, D.G.: Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Applied Statistics* **43** (1994), 429–467.
[8] Royston, P., Ambler, G., and Sauerbrei, W.: The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* **28** (1999), 964–974.
[9] Royston, P., and Sauerbrei, W.: Stability of multivariable fractional plynomial models with slection of variables and transformations: a bootstrap investigation. *Statistics in Medicine* **22** (2003), 639–659.
[10] Royston, P., and Sauerbrei, W.: *Multivariable Model-building*. John Wiley & Sons, Chichester, 2008.