ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 21, Number 1, June 2017
Available online at http://acutm.math.ut.ee

On estimation of insurance risk parameters by combining local regression and distribution fitting ideas

Meelis Käärik, Raul Kangro, and Liina Muru

Abstract. The problem of premium estimation is an essential part of the insurance mathematics. Often the problem is divided into two parts: estimation of claim number (or frequency) and the estimation of individual claim amounts (severities). In this paper, we will focus on the former. More precisely, we are looking for certain semiparametric dynamic regression type model to avoid the "price shock" issue of static classification. We apply locally the regression method, use local maximum likelihood estimation for the parameters of the model and cross-validation techniques to determine the optimal size of a neighborhood. A case study with real vehicle casco insurance dataset is included, the results obtained by proposed method are compared by the ones obtained by global regression and the classification and regression trees (C&RT) approach.

Introduction

The premium estimation problem can be divided into two parts: claim frequency and claim severity estimation. In this article we focus on the former and also try to avoid the caveats of some classical methods. In our previous works we have considered both problems, for claim severity estimation see, e.g., [7, 8, 6], whereas our more recent focus is on the claim frequency estimation [5, 9]. The current article can be considered as a follow-up to our previous works, where the classification and regression trees (C&RT) approach and the k-nearest neighbours approach were applied to estimate the claim frequency. For more details about the C&RT and k-nearest neighbours methods, see, e.g., [2] or [4].

Received September 30, 2016.
2010 *Mathematics Subject Classification.* 62P05; 62J05; 91B30.
Key words and phrases. Collective risk model; claim frequency estimation; local regression.
http://dx.doi.org/10.12697/ACUTM.2017.21.04

To further describe the background let us start with some well-known options that can be applied for the claim frequency estimation (or for premium estimation in general). As a first possible model, one can consider the naïve approach for classical collective risk model, where we first cluster the portfolio into homogeneous subportfolios, estimate the claim frequency (and severity) in each subportfolio, and finally estimate the total claim amount in each subportfolio and divide the expected claim amount (proportionally) between policies. Various different clustering and classification methods can be applied here (see, e.g., [4]), but the main issue of so-called "price shocks" still remains with any static classification. In short, the problem is based on the fact that small changes in client's data may result in large changes in premiums. The problem is especially large if the classification feature is a continuous variable like client's age or vehicle's age. Moreover, if a risk factor is a continuous variable, it is natural to assume that the risk premium is also continuously changing.

One possible class of models that does not suffer from the price-shock issue is the class of (generalized) linear models (see, e.g., [3]). On the other hand, in case of a "global" linear model it is difficult to find reasonable parametric forms for distribution parameters. It is not reasonable to assume that the distribution parameters of a policy are influenced by the behavior of clients with very different values of input parameters.

This reasoning motivates us to find a model that does not have the mentioned drawbacks of the classical models. Therefore we propose a combination of two widely used ideas:

- the k-nearest neighbours approach to find the neighborhood of similar risks for each policy (based on certain risk factors),
- the Poisson regression model applied to each risk taking into account the neighborhood corresponding to that risk.

Thus, the resulting model is a certain local regression model.

In this article we focus on the following issues.

- How to define a neighborhood in a meaningful way?
- What is the exact mathematical optimization problem behind the setup?
- What is the precision of prediction errors, or, when can we say that one method is really better than another?

The paper is organized as follows. In Section 1 we shall formulate the local regression model and provide the formulas for parameter estimation. In Section 2 we focus on the optimization problem and model comparison principles. A practical application is described in Section 3 and the obtained results are compared with the results of C&RT model obtained in [5]. Lastly, concluding comments are given in Section 4.

1. Local regression model

Let us start with the simplest case when we only have one risk factor (regressor).

1.1. The case with one regressor. Our estimation problem can be formulated as follows: we have a policy for which we would like to estimate the claims frequency. We only have one risk factor to use, and for the given policy, the value of that risk factor is denoted by x. We also have historical data of several policies and, based on certain rules (specified later), we find a neighborhood of policies that are close to our policy in the sense of given risk factor. Let $J(x)$ denote the set of indices of policies with regressor values in the neighborhood of value x, and for a policy with index i (or simply policy i) we denote

- x_i – the value of the regressor variable,
- t_i – number of days insured,
- n_i – number of claims.

Assuming the Poisson model for the claims frequency, the likelihood function can be written as

$$L_x(a,b) = \prod_{i\in J(x)} \frac{((a+b(x_i-x))t_i)^{n_i}}{n_i!} e^{-(a+b(x_i-x))t_i},$$

where a and b are the regression parameters.

The derivation of log-likelihood is straightforward and we obtain

$$l_x(a,b) = \sum_{i\in J(x)} n_i \ln((a+b(x_i-x))t_i) - \sum_{i\in J(x)} \ln(n_i!) - \sum_{i\in J(x)} (a+b(x_i-x))t_i.$$

Now, to find the maximum likelihood estimates, we first take the derivatives by parameters a and b

$$\frac{\partial l_x(a,b)}{\partial a} = \sum_{i\in J(x)} \frac{n_i}{a+b(x_i-x)} - \sum_{i\in J(x)} t_i,$$
$$\frac{\partial l_x(a,b)}{\partial b} = \sum_{i\in J(x)} \frac{n_i(x_i-x)}{a+b(x_i-x)} - \sum_{i\in J(x)} t_i(x_i-x),$$

from where the maximum likelihood estimates are obtained by solving

$$\begin{cases} \displaystyle\sum_{i\in J(x)} \frac{n_i}{a+b(x_i-x)} = \sum_{i\in J(x)} t_i, \\ \displaystyle\sum_{i\in J(x)} \frac{n_i(x_i-x)}{a+b(x_i-x)} = \sum_{i\in J(x)} t_i(x_i-x). \end{cases} \tag{1}$$

Remark 1. In case the policies in the neighborhood $J(x)$ are such that the risk factor has two distinct values (say, x_1 and x_2) only, there is an explicit solution

$$
\begin{aligned}
a + b(x_1 - x) &= \frac{n_1}{t_1}, \\
a + b(x_2 - x) &= \frac{n_2}{t_2},
\end{aligned}
$$

which yields

$$
\begin{aligned}
a &= \frac{n_2 t_1 (x_1 - x) + n_1 t_2 (x - x_2)}{t_1 t_2 (x_1 - x_2)}, \\
b &= \frac{n_1 t_2 - n_2 t_1}{t_1 t_2 (x_1 - x_2)}.
\end{aligned}
$$

1.2. The case with multiple regressors. In case of multiple regressors the situation becomes more complicated, but the general idea stays the same. Let us have m regressors, which means that the argument vector for a policy has the form $\mathbf{x} = (x_1, \dots, x_m)^T$ and, similarly, for the argument vector for a policy i we write $\mathbf{x}_i = (x_{1,i}, \dots, x_{m,i})^T$. We also need the following notation:

- $J(\mathbf{x})$ – the set of indices of policies with regressor values in the neighborhood of value $\mathbf{x}$,
- t_i – the number of days insured for policy i,
- n_i – the number of claims for policy i,
- $a, \mathbf{b} = (b_1, \dots, b_m)^T$ – the regression parameters.

Now, the likelihood function for the Poisson model can be expressed as

$$
L_{\mathbf{x}}(a, \mathbf{b}) = \prod_{i \in J(\mathbf{x})} \frac{((a + \mathbf{b}^T(\mathbf{x}_i - \mathbf{x}))t_i)^{n_i}}{n_i!} e^{-(a + \mathbf{b}^T(\mathbf{x}_i - \mathbf{x}))t_i}
$$

and the log-likelihood becomes

$$
\begin{aligned}
l_{\mathbf{x}}(a, \mathbf{b}) = & \sum_{i \in J(\mathbf{x})} n_i \ln((a + \mathbf{b}^T(\mathbf{x}_i - \mathbf{x}))t_i) \\
& - \sum_{i \in J(\mathbf{x})} \ln(n_i!) - \sum_{i \in J(\mathbf{x})} (a + \mathbf{b}^T(\mathbf{x}_i - \mathbf{x}))t_i
\end{aligned}
$$

Now, the derivatives by parameters have the forms

$$
\begin{aligned}
\frac{\partial l_{\mathbf{x}}(a, \mathbf{b})}{\partial a} &= \sum_{i \in J(\mathbf{x})} \frac{n_i}{a + \mathbf{b}^T(\mathbf{x}_i - \mathbf{x})} - \sum_{i \in J(\mathbf{x})} t_i, \\
\frac{\partial l_{\mathbf{x}}(a, \mathbf{b})}{\partial b_j} &= \sum_{i \in J(\mathbf{x})} \frac{n_i (x_{j,i} - x_j)}{a + \mathbf{b}^T(\mathbf{x}_i - \mathbf{x})} - \sum_{i \in J(\mathbf{x})} t_i (x_{j,i} - x_j), \quad j = 1, \dots, m,
\end{aligned}
$$

and the maximum likelihood estimates can be found from

$$\begin{cases} \sum_{i\in J(\mathbf{x})} \dfrac{n_i}{a+\mathbf{b}^T(\mathbf{x}_i-\mathbf{x})} = \sum_{i\in J(\mathbf{x})} t_i, \\ \sum_{i\in J(\mathbf{x})} \dfrac{n_i(x_{j,i}-x_j)}{a+\mathbf{b}^T(\mathbf{x}_i-\mathbf{x})} = \sum_{i\in J(\mathbf{x})} t_i(x_{j,i}-x_j), \quad j=1,\dots,m. \end{cases} \tag{2}$$

This system can be solved numerically, in our study we used the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [10, Section 10.7].

1.3. Determination of the neighborhood. Let us now consider the possible choices of determining the neighborhood of a policy. As we will apply the models in the cases of one or two regressors, we will also cover these setups here. In case of one regressor, the neighborhood can be fixed either based on the desired radius or the desired amount of policies (insurance years). Algorithmically, these approaches can be formulated as follows:

- fix the range of regressor values (say r), then all policies i are considered to be in the neighborhood of x whenever $|x_i - x| \le r$,
- fix the minimum neighborhood size k and increase the range r until sufficiently many policies (insurance years) are included.

We will use the combination of these two, so that a neighborhood will be determined both by minimum radius and minimum amount of insurance years. Notice that in this way we include both of the described approaches as a special case. The optimal values for minimal radius and the amount of insurance years included will be found by cross-validation.

In case of two regressors we consider similar approach:

- construct an elliptical neighborhood with given fixed radii r_1 and r_2 such that all policies are considered to be in the neighborhood of (x_1, x_2) if
$$\frac{(x_{1,i}-x_1)^2}{r_1^2} + \frac{(x_{2,i}-x_2)^2}{r_2^2} \le 1, \tag{3}$$
- fix the minimum neighborhood size k and increase the ranges r_1 and r_2 until sufficiently many policies (insurance years) are included.

While in the case of one single regressor the scaling of the regressor variable does not affect the result, in the case of two or more regressors the question of scaling is crucial. In two-regressor case one can restate the question as how to find the proper values of r_1 and r_2 or, equivalently, how to fix the ratio between r_1 and r_2.

Obviously, the simplest case would be to take $r_1 = r_2$ but then if the scales of the regressor variables are very different, one variable will dominate the other and the neighborhood is basically determined based on one regressor only. Also, simple rescaling of one regressor variable can result in completely different neighborhood selection, which is certainly an undesirable property. We propose three different scaling models here.

The first approach is based on the Euclidean distance and the regressors are scaled based on their standard deviations. In other words, we choose $r_1 = r \cdot \mathrm{sd}(X_1)$ and $r_2 = r \cdot \mathrm{sd}(X_2)$, where sd denotes the standard deviation, and we find the smallest value of r so that the neighborhood defined by (3) includes at least k insurance years. Notice that in this case the ellipse defining the neighborhood is always orthogonal to the axis.

The second approach is similar, but uses the Mahalanobis distance, which means that we also take into account the correlation between regressors X_1 and X_2. Thus the ellipse defining the neighborhood is no longer orthogonal to the axis, but shifted based on the correlation.

The third approach uses again the Euclidean distance, but the scaling is based on the span of the (local) regression line in a given neighborhood. More precisely, for a policy with regressor values x_1 and x_2, we first determine the optimal one-dimensional neighborhoods and the slopes of the one-dimensional local regression model, say b_1 and b_2. These values will be used to scale the regressors in the two-regressor model: we choose $r_1 = r/b_1$ and $r_2 = r/b_2$ and find the smallest value of r so that the neighborhood size is at least k.

2. Optimization problem and model comparison principles

We assume that the number of claims in each insured day for a policy with risk factors $\mathbf{x}$ is from the Poisson distribution with intensity $\lambda(\mathbf{x})$ and that for a given policy, the number of claims for different insured days are independent random variables. Each method of estimating the claim frequency corresponds to an estimator of that intensity, so we can think that each method produces a function $\hat{\lambda}(\mathbf{x})$. Clearly the goodness of a method should be related to how close is the prediction $\hat{\lambda}(\mathbf{x})$ to the actual intensity $\lambda(\mathbf{x})$ for all possible values of $\mathbf{x}$. As there are many ways to measure the distance between two functions, we should choose one that is the most relevant to our aims.

Let us assume that each policy corresponds to an iid realisation from a joint distribution of a triple of random variables (T, X, N), where T denotes the number of insured days, X denotes the vector of risk factors and N is the number of claims. Then a possible measure of closeness of the functions λ and $\hat{\lambda}$ (a loss function) is

$$D(\lambda, \hat{\lambda}) = \mathbf{E}[T\,(\lambda(X) - \hat{\lambda}(X))^2].$$

Note that since we predict the average number of claims per day, it is natural to multiply the squared prediction error by the number of insured days inside the loss function.

Since λ is not known, we cannot estimate the value of the loss function directly. We shall consider the following measure of the prediction error instead:

$$M(\hat{\lambda}) = \mathbf{E}[T\,(\hat{\lambda}(X) - \frac{N}{T})^2]. \tag{4}$$

Lemma 1. *Assume that the conditional distribution of N, given T and X, is the Poisson distribution with intensity $T\lambda(X)$. Then*

$$M(\hat{\lambda}) = D(\lambda, \hat{\lambda}) + \mathbf{E}\,\lambda(X).$$

Proof. By adding and subtracting $\lambda(X)$ inside the parentheses on the right hand side of (4) we get

$$\begin{aligned} M(\hat{\lambda}) &= \mathbf{E}[T\,(\hat{\lambda}(X) - \lambda(X))^2] + 2\mathbf{E}[(\hat{\lambda}(X) - \lambda(X))(T\lambda(X) - N)] \\ &\quad + \mathbf{E}[T\,(\lambda(X) - \frac{N}{T})^2] \\ &= D(\lambda, \hat{\lambda}) + 2\mathbf{E}[(\hat{\lambda}(X) - \lambda(X))(T\lambda(X) - N)] + \mathbf{E}[\frac{(T\lambda(X) - N)^2}{T}]. \end{aligned}$$

Using the well-known properties of conditional expectations and the assumption about the conditional distribution of N we have

$$\begin{aligned} \mathbf{E}[(\hat{\lambda}(X) - \lambda(X))(T\lambda(X) - N)] &= \mathbf{E}[\mathbf{E}[(\hat{\lambda}(X) - \lambda(X))(T\lambda(X) - N) \mid X, T]] \\ &= \mathbf{E}[(\hat{\lambda}(X) - \lambda(X))\mathbf{E}[(T\lambda(X) - N) \mid X, T]] \\ &= 0, \\ \mathbf{E}[\frac{(T\lambda(X) - N)^2}{T}] &= \mathbf{E}[\mathbf{E}[\frac{(T\lambda(X) - N)^2}{T} \mid X, T]] \\ &= \mathbf{E}[\frac{1}{T}\mathrm{Var}(N|X, T)] \\ &= \mathbf{E}[\lambda(X)]. \end{aligned}$$

Thus, indeed, the equality

$$M(\hat{\lambda}) = D(\lambda, \hat{\lambda}) + \mathbf{E}\,\lambda(X)$$

holds. □

Corollary 1. *If $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are two prediction functions and the assumption of Lemma 1 hold, then for any strictly increasing function $g : [0, \infty) \to \mathbb{R}$ we have*

$$g(M(\hat{\lambda}_1)) < g(M(\hat{\lambda}_2)) \Leftrightarrow D(\lambda, \hat{\lambda}_1) < D(\lambda, \hat{\lambda}_2).$$

This means that if we want to get the best predictor $\hat{\lambda}$ in a given class of prediction methods, we should aim to minimize an increasing function of $M(\hat{\lambda})$, for example $c\sqrt{M(\hat{\lambda})}$ for an appropriate scaling factor c.

If we have a sufficiently large sample of n policies, for which the claim numbers are known and which were not used in the process of constructing the prediction functions $\hat{\lambda}_1$ and $\hat{\lambda}_2$, we can use the estimates

$$\widehat{M(\hat{\lambda}_j)} = \frac{1}{n}\sum_{i=1}^{n} t_i(\hat{\lambda}_j(\mathbf{x}_i) - \frac{n_i}{t_i})^2, \ j = 1, 2,$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} t_i^2 \left((\hat{\lambda}_1(\mathbf{x}_i) - \frac{n_i}{t_i})^2 - (\hat{\lambda}_2(\mathbf{x}_i) - \frac{n_i}{t_i})^2\right)^2},$$

and since, by the construction, the estimates $\widehat{M(\hat{\lambda}_j)}$ are means of iid samples of the random variables $T(\lambda_j(X)+\frac{N}{T})^2$, they are asymptotically normal (see, for example, [1, p. 139]). This allows us to construct an approximate 95% confidence interval

$$(\widehat{M(\hat{\lambda}_1)} - \widehat{M(\hat{\lambda}_2)} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}, \widehat{M(\hat{\lambda}_1)} - \widehat{M(\hat{\lambda}_2)} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}) \tag{5}$$

for $M(\hat{\lambda}_1) - M(\hat{\lambda}_2)$. If this confidence interval does not contain 0, then we consider proved that one method is better than the other one.

3. Case study: Estonian casco insurance

In this section we will apply the proposed methods to real data.

3.1. Description of the data. The dataset used is obtained from an Estonian insurance company. The data covered 7 years of claim history, and the claims were classified by risk types (glass breakage risk, traffic accident risk, theft risk and more). Several important characteristics about the vehicle like the value, type, make, model and year of manufacture were available. Several characteristics about the owner of the vehicle (including sex, age and more) were also typically available.

Based on the earlier studies, the most important factors for claim frequency in case of traffic accident claims were the owner's age and the vehicle's age. For other risk types, the correlations to risk factors were weaker. Hence, in the numerical calculations we include owner's age and vehicle's age as arguments, and the traffic accident risk as the dependent variable. The owner's age ranged from 18 to 94 years and the vehicle's age ranged from 0 to 15 years.

3.2. Competing models. Based on the number of arguments we have two simple setups to estimate the frequency of traffic accident claims:

- **Setup 1**: age of the owner as the argument.
- **Setup 2**: age of the owner and age of the vehicle as arguments.

The competing models (for both setups) are:

- C&RT/Poisson model,
- local regression model (with minimal radius and neighborhood size determined by cross-validation),
- "global" regression model.

In case of Setup 2, there are actually three different local regression models, based on different scaling methods of the regressors (see Section 1.3).

For comparison of different models, the dataset was divided into two parts: training data and test data. All proposed models were calibrated on training data and then the model with same parameters was applied to test data. Optimal values for neighborhood size and minimal radius were determined by 10-fold cross-validation, the regression parameters were found by solving the MLE equations (1) or (2). The "goodness" of a model was measured by the error characteristic

$$e = \sqrt{\frac{\sum_{i=1}^{n} t_i (365\frac{n_i}{t_i} - 365\hat{\lambda}(\mathbf{x}_i))^2}{\sum_{i=1}^{n} t_i}},$$

where n is the number of policies in test data, t_i is the number of days insured for policy i, n_i is the actual number of claims for policy i, and $\hat{\lambda}(\mathbf{x}_i)$ is the predicted daily frequency of claims for policy i. Note that, using the notation of Section 2, we have

$$e = 365\sqrt{\frac{n}{\sum_{i=1}^{n} t_i}\widehat{M(\hat{\lambda})}},$$

so we are comparing different prediction functions by the method proposed in Section 2. This form of the goodness measure was chosen for the compatibility with the earlier studies.

3.3. Results, one regressor. To determine the optimal neighborhood size and minimal radius 10-fold cross-validation was used. The results for age of owner as the regressor variable and different choices of radius r and neighborhood size k are shown in Figure 1. As one can see from the figure, the optimum is reached when minimal radius is 7 years and neighborhood size is 500 insurance years.

Similar calculations are carried out for the model where age of the vehicle is the regressor variable. Notice that these calculations are only needed for the two-regressor model of third type, where the slopes of one-dimensional local regression models are used for scaling. The results are presented in Figure 2. As one can see from the figure, the minimal model error is obtained when minimal radius is 10 years and neighborhood size is 3200 insurance years. One can also notice that the choice of neighborhood size barely affects the outcome, which hints that the age of the vehicle is not very informative in the sense of determining the claim frequency.

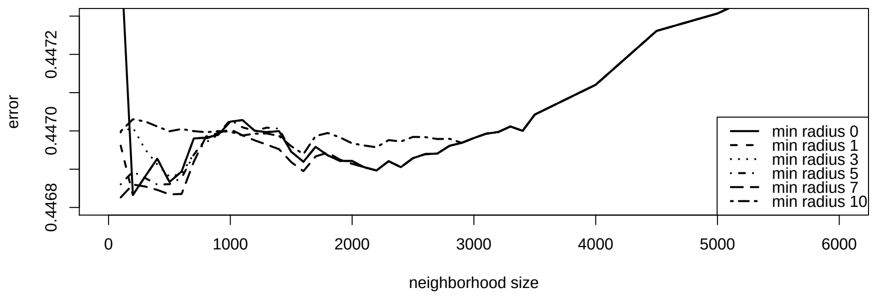

FIGURE 1. Cross-validation errors using age of the owner as the regressor.

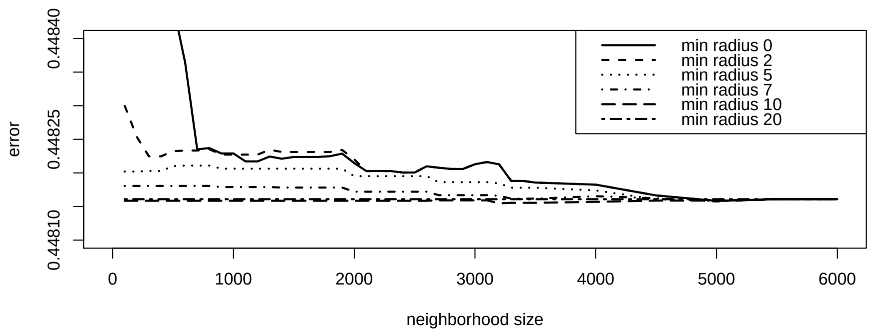

FIGURE 2. Cross-validation errors using age of the vehicle as the regressor.

Now, the regressor model with age of the owner as the regressor with above obtained parameters is compared with the results from C&RT model [5] and with the "global" Poisson regression model. The results are shown in the following table. We see that our proposed local regression model was the best in the sense of the model error on test data.

Model	**Error**
C&RT	0.67204
Local regression	0.65723
"Global" regression	0.65747

TABLE 1. Comparison of models, Setup 1.

The pairwise construction of confidence intervals (5) for given methods showed that the difference between either regression method and C&RT was significant, i.e., we can say that the regression methods work better than C&RT in this setup. On the other hand, the difference between results of the two regression methods was not significant (by the same criterion).

3.4. Results, two regressors. Similarly to the one-regressor model, the optimal neighborhood size for two-regressor model is determined by 10-fold cross-validation. In all cases the regressor variables were the age of the owner (X_1) and the age of the vehicle (X_2), and the following three different models were proposed based on the scaling of individual regressor variables:

- Model 1: Euclidean distance is used and both regressor variables are scaled by their variance, i.e., in Formula (3) we choose $r_1 = r \cdot \mathrm{sd}(X_1)$ and $r_2 = r \cdot \mathrm{sd}(X_2)$;
- Model 2: Mahalanobis distance is used, i.e., besides scaling by variance also the correlation between regressors is taken into account;
- Model 3: scaling is based on slopes of one-regressor models obtained in previous subsection, i.e., in Formula (3) we choose $r_1 = r/b_1$ and $r_2 = r/b_2$, where b_1 is the slope of the one-regressor model with age of the owner and b_2 is the slope of the one-regressor model with age of the vehicle.

The covariance matrix and correlation matrix for the regressor variables are the following:

$$\mathrm{Cov}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 151.6 & -3.98 \\ -3.98 & 7.57 \end{pmatrix} \text{ and } \mathrm{Corr}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & -0.117 \\ -0.117 & 1 \end{pmatrix}.$$

As is seen from the covariance matrix, the variability of the age of the owner is much bigger than the variability of the age of the vehicle, which means that the ellipse (3) defining the neighborhood allows more variability within the age of the owner. Notice also that there is only mild correlation between regressor variables, which indicates that the results for Models 1 and 2 are expected to be quite similar.

The cross-validation results are shown in Figure 3. Based on cross-validation, the optimal values for neighborhood size k are:

- $k = 900$ for Model 1,
- $k = 1000$ for Model 2,
- $k = 2300$ for Model 3.

Finally, all these models are applied to test data, and the results of proposed models together with the C&RT model and with the "global" Poisson regression model are shown in the following table.

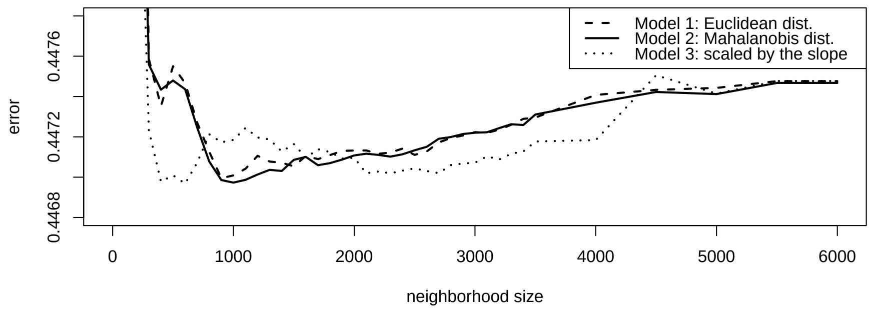

FIGURE 3. Cross-validation results for different 2-regressor models.

Model	**Error**
C&RT	0.67202
Local regression, model 1 (Euclidean dist.)	0.65720
Local regression, model 2 (Mahalanobis dist.)	0.65714
Local regression, model 3 (scaling by slope)	0.65744
"Global" regression	0.65777

TABLE 2. Comparison of models, Setup 2.

The pairwise construction of confidence intervals (5) gave similar results to one argument setup: the performance of all the proposed regression methods was significantly better than the performance of C&RT (in the sense of (5)), but none of the differences between different regression methods was significant.

4. Summary

In the paper we proposed various ideas that can be used for implementing a local regression model for premium estimation. Although the theory was presented only in the context of estimating the claim frequency for different risk factors, the ideas can be easily extended also to the estimation of the expected claim severity (and thus to the estimation of the expected losses). We also provide clear principles for model selection and comparison. These principles are quite general and valid for a broader class optimization problems. The empirical study confirms that local regression models may have a clear advantages over the classical methods of dividing the insurance portfolio into homogeneous classes with similar risk factors. The competitive advantages/disadvantages of various implementation setups proposed in the paper clearly deserve further research.

Acknowledgments

The research was supported by institutional research funding IUT34-5 of the Estonian Ministry of Education and Research. The authors thank editors and anonymous referees for their helpful and constructive comments and suggestions.

References

[1] G. Blom, *Probability and Statistics: Theory and Applications*, Springer-Verlag, New York, 1989.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Belmont, Wadsworth, 1984.

[3] P. de Jong and Z. H.. Heller, *Generalized Linear Models for Insurance Data*, Cambridge University Press, Cambridge, 2008.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition*, Springer, New York, 2009.

[5] M. Käärik and A. Kaasik, *On premium estimation using the C&RT/Poisson model and its extensions*, Lithuanian J. Statist. **51**(1) (2012), 5–13.

[6] M. Käärik and H. Kadarik, *Statistical inference with the limited expected value function*, in: Multivariate Statistics: Theory and Applications, World Scientific, 2013, pp. 99–111.

[7] M. Käärik and M. Umbleja, *On claim size fitting and rough estimation of risk premiums based on Estonian traffic insurance example*, Internat. J. Math. Models Methods Appl. Sci. **5**(1) (2011), 17–24.

[8] M. Käärik and A. Žegulova, *On estimation of loss distributions and risk measures*, Acta Comment. Univ. Tartu. Math. **16**(1) (2012), 53–67.

[9] K. Pärna, R. Kangro, A. Kaasik, and M. Möls, *K-nearest neighbors as pricing tool in insurance: a comparative study*, in: Multivariate Statistics: Theory and Applications, World Scientific, 2013, pp. 130–140.

[10] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, and P. B. Kramer, *Numerical Recipes: the Art of Scientific Computing*, Cambridge University Press, Cambridge, 1987.

Institute of Mathematics and Statistics, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia

E-mail address: meelis.kaarik@ut.ee

E-mail address: raul.kangro@ut.ee

E-mail address: liinamur@gmail.com