J. Appl. Math. & Informatics Vol. **32***(2014), No. 3 - 4, pp. 343 - 357*
http://dx.doi.org/10.14317/jami.2014.343

FUZZY REGRESSION TOWARDS A GENERAL INSURANCE APPLICATION†

JOSEPH H. T. KIM AND JOOCHEOL KIM*

Abstract. In many non-life insurance applications past data are given in a form known as the run-off triangle. Smoothing such data using parametric crisp regression models has long served as the basis of estimating future claim amounts and the reserves set aside to protect the insurer from future losses. In this article a fuzzy counterpart of the Hoerl curve, a well-known claim reserving regression model, is proposed to analyze the past claim data and to determine the reserves. The fuzzy Hoerl curve is more flexible and general than the one considered in the previous fuzzy literature in that it includes a categorical variable with multiple explanatory variables, which requires the development of the fuzzy analysis of covariance, or fuzzy ANCOVA. Using an actual insurance run-off claim data we show that the suggested fuzzy Hoerl curve based on the fuzzy ANCOVA gives reasonable claim reserves without stringent assumptions needed for the traditional regression approach in claim reserving.

AMS Mathematics Subject Classification : 62A86, 62P20, 91B30.
Key words and phrases : Fuzzy regression, chain ladder, claim reserving.

1. Introduction

In non-life insurance applications determining the evolution of the future claims is an important consideration for insurance companies. The estimated amount of future claims then forms a basis for the reserve which must be set aside to protect the insurer from future losses. In the current article we focus on non-life insurance contracts, such as the auto and medical insurance, and attempt to find the fair reserve using the fuzzy regression methodology.

Received October 4, 2013. Revised December 15, 2013. Accepted February 10, 2014.
*Corresponding author.
†This work was supported by the research grant of Yonsei University. J.H.T. Kim is grateful for the support from Basic Science Research Program of the National Research Foundation of Korea (NRF-2012R1A1A1043439)

The traditional claim reserving approach in the insurance literature typically takes the following sequential steps to determine the reserve amount to be held by the insurer for an insurance portfolio. First, the claim trend is estimated from the past claim data using some standard regression models. Second, assuming that the future claim pattern would emerge in a similar fashion as observed in the past, future claims are projected based on the estimated regression model. This step also allows that the stochastic characteristic of the future claims can be captured by the perturbation term of the regression error term. Finally, using the predicted claim amounts, the reserve of the portfolio is determined by the difference between the predicted ultimate future claim amount and the (known) current claim amount.

While this crisp approach can capture some stochastic aspects of future uncertainty, its adoption of the standard regression models is criticized on several bases. For example, the number of data to fit the regression model is typically of small size, which could lead to inadequate statistical analyses. Also, more importantly, the set of error assumptions required for regression analyses, such as the independence among the perturbation terms, is easily violated under the crisp approach. This may seriously distort the credibility of the predicted future claims as well as the degree of its uncertainty.

In light of these shortcomings, other alternatives and generalizations of insurance claim reserving methods have been proposed in the literature; see, e.g., [2] for a survey of various reserving schemes. Among these [8] offers an alternative reserving method based on a fuzzy theory. In its original paper, [8] utilizes the simple fuzzy linear regression of [3] on the link ratio[1] on the log-transformed past data. The reserves obtained from this fuzzy regression is reported to perform well compared to the traditional crisp reserving approach, with less stringent assumptions on the error terms of the ordinary least square regression method. [8] also provides a concise survey for other insurance applications of the fuzzy theory.

Our contribution in this paper is twofold. First, we extend the fuzzy claim reserving method of [8] where the simple fuzzy regression is adopted ignoring the cohort (that is, calendar year) effect in the claim data. We employ a more general parametric called the Hoerl curve which accommodates the cohort effect as well as the development periods; see, e.g., [2], [11] and [4]. The use of the Hoerl curve, however, calls for a statistical analysis known as the analysis of covariance, or ANCOVA, a combination of the linear regression and the analysis of variance. Therefore our second contribution is this paper is a development of the fuzzy ANCOVA model.

The present article is organized as follows. In Section 2 some backgrounds on the fuzzy numbers and regression are presented. Section 3 explains how the crisp regression method is used for the traditional claim reserving in insurance applications. In particular, the Hoerl curve is introduced as a flexible parametric

[1]See the next section for details.

model, which is an ANCOVA model, a blending of the linear regression model and the analysis of variance (ANOVA). In Section 4, the fuzzy counterpart of the crisp Hoerl curve is proposed along with the fuzzy ANCOVA procedure. The resulting fuzzy reserves are also calculated for the working data. Throughout the paper we use an actual insurance data retrieved from [4] for numerical illustrations.

2. Fuzzy numbers and fuzzy regression

2.1. Fuzzy numbers. A fuzzy number (FN) is a fuzzy subset $\tilde{a}$ defined over real numbers. Among different choices of FNs we focus on Triangular FNs (TFNs) for its practicality and mathematical tractability. A TFN is defined as $\tilde{a} = (a, l_a\, r_a)$ where a is the center (or core) and the latter two stand for the left and right spreads, respectively. A characterization of such a FN can be made explicit via its membership function

$$\mu_{\tilde{a}}(x) = \begin{cases} \frac{x-a+l_a}{l_a} & a - l_a < x \le a, \\ \frac{a+r_a-x}{r_a} & a < x \le a + r_a, \\ 0 & \text{otherwise.} \end{cases}$$

or, alternatively, by its α-cuts:

$$a_\alpha = [\underline{a}(\alpha), \overline{a}(\alpha)] = [a - l_a(1-\alpha),\, a + r_a(1-\alpha)] \tag{1}$$

2.2. Fuzzy regression. Consider an n-variate crisp function $y = f(\theta_1, ..., \theta_n)$. In the regression context θ_i is the ith regression coefficient. If $\theta_1, ..., \theta_n$ are *not* crisp numbers but FNs $\tilde{a}_1, ..., \tilde{a}_n$, we have

$$\tilde{b} = f(\tilde{a}_1, ..., \tilde{a}_n) \tag{2}$$

If we restrict $f(\cdot)$ to be linear so that $\tilde{b} = f(\tilde{a}_1, ..., \tilde{a}_n) = \sum_{i=1}^n \tilde{a}_i x_i$, where symbol x_i has been deliberately chosen to relate to regression models, the resulting $\tilde{b}$ is again a TFN with the three elements given by

$$b = \sum_{i=1}^n a_i x_i, \quad l_b = \sum_{i=1, x_i \ge 0}^n l_{a_i}|x_i| + \sum_{i=1, x_i < 0}^n r_{a_i}|x_i|,$$

and

$$r_b = \sum_{i=1, x_i \ge 0}^n r_{a_i}|x_i| + \sum_{i=1, x_i < 0}^n l_{a_i}|x_i| \tag{3}$$

In fact, for the linear functional case, we can not only obtain $\tilde{b}$, but the closed expression for the $\alpha-$cuts of $\tilde{b}$ as well. Let us suppose without loss of generality that f is increasing in the first $m \le n$ variables (i.e., $\theta_1, ..., \theta_m$) and decreasing in the remaining variables (i.e., $\theta_{m+1}, ..., \theta_n$), $\tilde{b}$'s α-cuts are then simply

$$\tilde{b}_\alpha = \big[f(\underline{a}_1(\alpha), ..., \underline{a}_m(\alpha), \overline{a}_{m+1}(\alpha), ..., \overline{a}_n(\alpha)), \\ f(\overline{a}_1(\alpha), ..., \overline{a}_m(\alpha), \underline{a}_{m+1}(\alpha), ..., \underline{a}_n(\alpha))\big] \tag{4}$$

Now we describe the fuzzy regression (FR) of [3], an extension of [10]. The FR to be introduced here is a natural applications of the linear function result explained above. Consider a sample of size n with m explanatory variables. The FR is then stated as

$$\tilde{Y}_j = \tilde{a}_0 + \tilde{a}_1 X_{1j} + ... + \tilde{a}_m X_{mj} \tag{5}$$

where the coefficients are fuzzy and the explanatory variables are crisp. Assuming the TFN for all FNs, this has a nice solution for $\tilde{Y}_j = (Y_j, l_{Y_j}, r_{Y_j})$ as before:

$$Y_j = a_0 + \sum_i^m a_i X_{ij}, \quad l_{Y_j} = l_{a_0} + \sum_{x_{ij} \geq 0} l_{a_i}|X_{ij}| + \sum_{x_{ij}<0} r_{a_i}|X_{ij}| \tag{6}$$

$$\text{and} \qquad r_{Y_j} = r_{a_0} + \sum_{x_{ij} \geq 0} r_{a_i}|X_{ij}| + \sum_{x_{ij}<0} l_{a_i}|X_{ij}| \tag{7}$$

To estimate FNs $\tilde{a}_0, \tilde{a}_1, ..., \tilde{a}_m$ we take the following two steps:

(1) The cores of $\tilde{a}_0, \tilde{a}_1, ..., \tilde{a}_m$ are estimated using the ordinary least squares (OLS) regression method. The estimated cores are denoted $\hat{a}_0, \hat{a}_1, ..., \hat{a}_m$

(2) The left and right spreads of $\tilde{a}_0, \tilde{a}_1, ..., \tilde{a}_m$ are estimated from the following optimization, for a given minimum accomplishment level α^*:

Minimize: $\sum_{j=1}^{n}\sum_{i=0}^{m} l_{a_i}|X_{ij}| + \sum_{j=1}^{n}\sum_{i=0}^{m} r_{a_i}|X_{ij}|$

Subject to

$$\hat{a}_0 + \sum_{i=1}^{m} \hat{a}_i X_{ij} - \left[l_{a_0} + \sum_{x_{ij} \geq 0} l_{a_i}|X_{ij}| + \sum_{x_{ij}<0} r_{a_i}|X_{ij}| \right] (1-\alpha^*) \leq \underline{Y_j}$$

$$\hat{a}_0 + \sum_{i=1}^{m} \hat{a}_i X_{ij} + \left[r_{a_0} + \sum_{x_{ij} \geq 0} r_{a_i}|X_{ij}| + \sum_{x_{ij}<0} l_{a_i}|X_{ij}| \right] (1-\alpha^*) \geq \overline{Y_j}$$

$$j = 1, ..., n, \quad l_{a_i}, r_{a_i} \geq 0, i = 0, ..., m$$

Here the lower and upper values of the response, $\underline{Y_j}$ and $\overline{Y_j}$, are set to be the minimum and maximum possible value of Y_j, respectively. If there are repeated observations at level j one may simply take $\underline{Y_j}$ (or $\overline{Y_j}$) to be the maximum (or minimum) among the observations at level j. Otherwise, these are appropriately selected by the experimenter. Hence we can obtain $\tilde{a}, \tilde{b}$ using the optimization algorithm above at any given level $\alpha^* \in [0, 1]$. We will use this algorithm to fit the actual data later.

When $f(\theta_1, ..., \theta_n)$ is a non-linear function, its FN version (2) is not suitable for TFN case as the left side $\tilde{b}$ is not in general a TFN even though $\tilde{a}_1, ..., \tilde{a}_n$ are. One suggestion to resolve this conflict is to use the first-order Talyor expansion to approximate $\tilde{b}$ with a TFN $\tilde{b}'$; see [1]. Assume again that $f(\theta_1, ..., \theta_n)$ is increasing in the first $m \leq n$ variables and decreasing in the remaining variables.

Then the linear (or the first-order Talyor) approximation $\tilde{b}' = (b, l_b, r_b)$ is given by

$$b = f(a_1, ..., a_n);\ \ l_b = \sum_{i=1}^{m} \frac{\partial f}{\partial \theta_i} l_{a_i} - \sum_{i=m+1}^{n} \frac{\partial f}{\partial \theta_i} r_{a_i};\ \ r_b = \sum_{i=1}^{m} \frac{\partial f}{\partial \theta_i} r_{a_i} - \sum_{i=m+1}^{n} \frac{\partial f}{\partial \theta_i} l_{a_i} \quad (8)$$

For the linear programming in the fuzzy contexts, refer to, for example, [5] and [6].

3. Claim reserving method: Traditional crisp approach

In this section we introduce the concept to insurance claim reserving using the traditional crisp approach that is based on the regression analysis.

3.1. Background. For general insurance business, contracts, insurance contracts may have a long period to settle claims due to, e.g., legal processes. For example, consider the accidents occurred in 2013. For these accidents, the insurer makes claim payments not just in 2013, but also in the subsequent years 2014, 2015, and so on. Hence the insurance premium collected in 2013 must be large enough to cover the claims arisen from multi-years from 2013. In our example, the claim payment made in each year (2013, 2014, ...) is called the incremental loss belonging to *accident year* 2013, and each year after 2013 is termed the development year. Because of this time lag effect, one would expect the incremental loss eventually decreases over time (development years), converging to zero.

Clearly, the accident year, which refers to the origin time of a given loss, is different from the *calendar year*, the usual year we use everyday. It is a standard practice that the total claim payments made in any calendar year are split and attributed to each accident year. For example, the total claim payments made in calendar year 2013 covers not just the claims occurred in the current accident year 2013, but also claims belonging to past accident years 2012, 2011, 2010, and so on. Understanding how much proportion of the current calendar year payment belongs to each accident year is important for the insurer as it provides the basis of the premium for insurance contracts as aforementioned.

Due to this complication, the historical insurance claim data is typically presented in a so-called run-off triangle looking like Table 1, which is an actual data retrieved from [4] (Chapter 10). Each row represents the accident year and each columns stands for the development year (period). In the table, $C_{i,j}$ is the incremental loss payment made in development year j originated from accident year i. Consequently $i + j = n$ is the calendar year where $C_{i,j}$ is made, and thus $\sum_{i+j=n} C_{i,j}, (\forall i, j)$ stands for the total payments of the insurer in calendar year n. For notational simplicity, it is customary for i to take values $1, ..., n$ so that $i = 1$ corresponds to the initial accident year reported in data, and $i = n$ to the latest calendar year observed. The empty elements in the table are future loss values to be predicted. The main task of run-off triangle analysis insurance claim reserving is to fill the table with a suitably predicted numbers from a model.

TABLE 1. Run-off table of incremental loss $C_{i,j}$ data

	Development yr (j)							
Accid yr (i)	1	2	3	4	5	6	7	8
1	101	153	52	17	14	3	4	1
2	99	121	76	32	10	3	1	
3	110	182	80	20	21	2		
4	160	197	82	38	19			
5	161	254	85	46				
6	185	201	86					
7	178	261						
8	168							

TABLE 2. Run-off table of cumulative loss $Z_{i,j}$ data

	Development yr (j)							
Accid yr (i)	1	2	3	4	5	6	7	8
1	101	254	306	323	337	340	344	345
2	99	220	296	328	338	341	342	
3	110	292	372	392	413	415		
4	160	357	439	477	496			
5	161	415	500	546				
6	185	386	472					
7	178	439						
8	168							

In addition to the incremental loss $C_{i,j}$, the standard claim reserving requires several related quantities:

- The cumulative loss:

$$Z_{i,j} := \sum_{k=1}^{j} C_{i,k}, \quad i = 1, ..., n\,;\ j = 1, 1, ..., n - i \tag{9}$$

  The run-off table of $Z_{i,j}$ for the data is presented in Table 2
- The link ratio:

$$r_{i,j} = \frac{Z_{i,j+1}}{Z_{i,j}}, \quad i = 1, ..., n-1\,;\ j = 1, ..., n - i - 1 \tag{10}$$

  Note that $r_{i,j} \geq 1$ always, and eventually converges to 1 as j gets larger
- If multiplied successively, the link ratio gives the so-called future projection rate. We define the future projection rate regarding accident year i from the development period j to period s, with $j < s$, as

$$f_{j,s}^{(i)} = \prod_{h=j}^{s-1} r_{i,h} = \frac{Z_{i,s}}{Z_{i,j}} \tag{11}$$

Often parametric models for $C_{i,j}$ are selected so that the right side of (11) is independent of i; this is the case when $C_{i,j}$ has a multiplicative term involving i only.

3.2. Finding past claim trend using ANCOVA.

3.2.1. Modeling claim trend with regression. In order to predict the future claim we first need to find the past claim trend. Traditionally insurers used the past values of $C_{i,j}$, $Z_{i,j}$ or $r_{i,j}$ to model the past claim trend. Various parametric models have been suggested in the literature. Some use $r_{i,j}$ values to model the past claim trend, while others use $C_{i,j}$ or $Z_{i,j}$; see, e.g., [2] for a survey. In [8] the log-transformed link ratio is regressed on the development year, ignoring the calendar year effect, to yield a simple linear FR, the idea motivated from the crisp approach of [9]. That is, the regression equation

$$\log(r_j - 1) = a + b \log(j + 1) \tag{12}$$

is analyzed using the OLS.

In the present article, we consider a more general parametric model called the Hoerl curve, as discussed in [2], [11] and [4] (Chapter 10). In the Hoerl curve the incremental loss, rather than the link ratio, is modeled directly:

$$C_{i,j} = \exp(c + \alpha_i) \exp(\beta \log j + \gamma j) \tag{13}$$

After taking logarithm on both sides, one arrives at the following linear equation

$$\log(C_{i,j}) = c + \alpha_i + \beta \log(j) + \gamma j \tag{14}$$

If perturbation terms are added, (14) can be analyzed in the linear regression framework with suitably estimated parameters for the mean responses. From the regression perspective, the Hoerl curve after log-transform (14) has several advantages over the Sherman's model (12) considered in [8]. First, the Hoerl curve is more flexible as it allows two explanatory variables, leading to a multiple linear regression rather than the simple one in [8]. The power of explaining the response variable therefore should be better in general. Second, unlike the Sherman's model, the Hoerl curve takes the calendar year (or cohort) effect into account. This is a desirable aspect of any claim reserving model as the same cohort could have a common characteristic shared over time, such as the inflation factor. Overall, the Hoerl curve provides a more realistic parametric model with additional parameters over the Sherman's model. The additional parameters however change the structure of the regression model, leading to a model commonly known as the analysis of covariance (ANCOVA) in the statistical literature.

3.2.2. ANCOVA. In the statistical literature, the ANCOVA essentially blends the linear regression model and the analysis of variance (ANOVA). The model considered in ANCOVA is typically a multiple regression analysis in which there is at least one quantitative and one categorical explanatory variable. Usually the discrete categorical variable stands for different groups or factors (e.g., different

types of treatments, genders), and the quantitative variables are control variables which are included to improve power. Restricting on the case where there is one categorical variable, and no interaction between the categorical variable and other m quantitative variables, the regression equation looks like:

$$Y_{ij} = \mu + \tau_i + a_1 X_{ij}^1 + ... + a_m X_{ij}^m \tag{15}$$

where $\mu + \tau_i$ stands for the effect of the ith group or treatment in the categorial variable, and X^k is the kth quantitative explanatory variable. The model in (15) thus relates the response variable with both categorical and quantitative variables. Note however that the intercept is the only coefficient that varies over different groups; other coefficients are common for all the groups. This indicates that after neutralizing the group effect by adjusting the intercept terms, all the data share the same regression model. An alternative modeling approach in the presence of the categorial variable is to simply treat it as another explanatory variable, in which case the standard regression can be carried out with no additional difficulties. It is noted however that such an approach is different from the model stated in (15) as, in this case, the coefficient of the categorical variable extracts the linear trend in the the categorial variable assuming its continuity. Hence it cannot capture the qualitative distinction among different groups. See standard statistics texts, e.g., [7], for further details on ANCOVA analyses. In light of this, we see that the regression equation (14) is a special case of the model stated in (15), and thus suitable for ANCOVA analysis. In particular, in (14), index i or the whole intercept term $c + \alpha_i$ explains the effect of different calendar year in the run-off triangle. Other variabilities of the data is explained by the development year j and its function $\log j$. So, after adjusting the pre-existing differences in calendar year effect, all claims should evolve in the same fashion as a function of the development year. Using the standard statistical package, we obtain the parameters for the working data as

$$\begin{aligned}(\hat{c} + \hat{\alpha}_1, ..., \hat{c} + \hat{\alpha}_8) &= (6.163, 5.951, 6.079, 6.445, 6.542, 6.519, 6.706, 6.442) \\ \hat{\beta} &= 1.85624; \quad \hat{\gamma} = -1.31755\end{aligned} \tag{16}$$

The ANCOVA is a widely used analysis method in statistics but its fuzzy counterpart is rarely studied in the literature. In the next section we propose a fuzzy ANCOVA method based on the FR of [3], and use it to predict the future claims and estimate the fair reserves for an insurance portfolio. In the passing, we provide the estimated parameters of the alternative model where the categorical variable is treated as a explanatory variable. The estimated parameters are $\hat{\mu} = 5.931, \hat{\beta} = 1.880, \hat{\gamma} = -1.325$, and the coefficient for the accident year is 0.0982, which is positive and always leads to a higher log incremental loss for higher calendar year, contradicting the intercept pattern in (16) where some later calendar years have smaller intercepts. In the sequel, we focus on the ANCOVA model only.

3.3. Predicting future claim evolution. The estimated regression models in (14) basically smooths the past claim trend. In the insurance claim reserving it is the convention to assume that the future claim pattern would emerge in a similar fashion as observed in the past, and thus future claims are projected based on the estimated regression model. For this, one first needs to back-transform to recover the original quantities. The link ratio is then, from (10) and the Hoerl curve (13),

$$r_{i,j} = \frac{Z_{i,j+1}}{Z_{i,j}} = \frac{\sum_{k=1}^{j+1} C_{i,k}}{\sum_{k=1}^{j} C_{i,k}} = \frac{\sum_{k=1}^{j+1} \exp(\beta \log k + \gamma k)}{\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)}, \tag{17}$$

which simplifies things as this is independent of i. Hence, the future projection rate (11) takes a simple form without index i (so the superscript is omitted):

$$f_{j,s} = \frac{\sum_{k=1}^{s} \exp(\beta \log k + \gamma k)}{\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)} \tag{18}$$

The ultimate future cumulative loss $Z_{i,n}$, defined as the last column of the run-off table of $Z_{i,j}$, is then estimated as the product of the most recent cumulative loss and future projection factor for the future time horizon:

$$Z_{i,n} = Z_{i,n-i+1} f_{n-i+1,n} \tag{19}$$

Finally, the reserve for accident year i, determined in calendar year n, is defined as the difference between the ultimate future cumulative loss and the current cumulative loss

$$R_i = Z_{i,n} - Z_{i,n-i+1} = Z_{i,n-i+1} f_{n-i+1,n} - Z_{i,n-i+1} \tag{20}$$

The reserve is understood most easily using Table 2. Here R_i is the difference between the last column value of row i after the table has been fully filled and the latest value of the triangle in row i. The total reserve then is simply the sum of the reserves of all accident years

$$R = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} (Z_{i,n-i+1} f_{n-i+1,n} - Z_{i,n-i+1}) \tag{21}$$

For the numerical data, the crisp reserve results are provided in the top panel of Table 3, which will be further discussed in Section 4.

4. Claim reserving method using fuzzy regression

4.1. Motivation. We have shown how the classical (crisp) claim reserving approach uses a linear regression model to smooth past claim data as in (14) and predict the projection rates. However the assumptions underlying the standard regression model are criticized on several grounds as briefly mentioned in Introduction. Specifically, the error terms added in (14) are neither independent nor identically distributed. This is because the past claims are unlikely to be uncorrelated. Also, when the data size is small, as is often the case in insurance applications, statistical analyses may not give meaning ful conclusions. In this

section, we now look at the claim reserving with fuzzy regression method as an alternative solution to overcome the difficulties of the standard approach.

4.2. Fuzzy ANCOVA. Our fuzzy ANCOVA adapts the fuzzy regression method proposed by [3], of which procedure is described in Section 2.2. As before we assume that the regression coefficients are fuzzy while the explanatory variables are crisp. To begin with, we may consider the fuzzy counterpart of (15):

$$\tilde{Y}_{ij} = \tilde{\mu} + \tilde{\tau}_i + \tilde{a}_1 X_{ij}^1 + ... + \tilde{a}_m X_{ij}^m \tag{22}$$

which is different from (5), and clearly the FR approach in Section2 is not directly applicable. To tackle this problem, we first recall that, in the crisp ANCOVA model (15), the data should share the same regression model after an adjustment for preexisting differences in nonequivalent groups. Another way to look at this is to rearrange (22) to get

$$\tilde{Y}_{ij} - \tilde{\mu} - \tilde{\tau}_i = \tilde{a}_1 X_{ij}^1 + ... + \tilde{a}_m X_{ij}^m \tag{23}$$

The left side is then the response variable net of the group or treatment effect, making the right side no longer affected by index i. Consequently, we put the left side $\tilde{Y}_j^* = \tilde{Y}_{ij} - \tilde{\mu} - \tilde{\tau}_i$, omitting i, and can further express (23) as

$$\tilde{Y}_j^* = \tilde{a}_1 X_j^1 + ... + \tilde{a}_m X_j^m \tag{24}$$

which is equivalent to the FR in (5) without the intercept term. If there are w different groups (treatments) and n_i observations for $i = 1, ..., w$, the index j runs over $j = 1, 2, ..., \sum_{i=1}^{w} n_i$. As the solution of the FR in (24) can be readily available from Section 2, the remaining task is to determine the intercept $\tilde{\mu} + \tilde{\tau}_i$, for each i, so that the final FN response variable can be obtained from

$$\tilde{Y}_{ij} = \tilde{Y}_j^* + \tilde{\mu} + \tilde{\tau}_i \tag{25}$$

Essentially the challenge lies in estimating the intercept FN separately, in the presence of the other FN coefficients in the FR model (5). In theory the intercept FN $\tilde{\mu} + \tilde{\tau}_i$ can vary in their center values as well as the spreads over different i. If we assume however that the categorical variable is a description of a qualitative type or classes, as is the case for most applications, and thus cannot be fuzzy by nature, we could argue that both $\tilde{\mu}$ and $\tilde{\tau}_i$ be crisp numbers obtained from the OLS in ANCOVA. We believe this solution is consistent with the spirit of the ANCOVA because term $\tilde{\mu} + \tilde{a}_i$ is the only source representing the categorial variable in the model (22), and this source should not be fuzzy by nature as it categorizes, e.g., different genders (groups), different types of treatments, or different calendar years in our case.

To summarize, the estimation procedure of the fuzzy ANCOVA analysis for (22) is as follows:

Step 1: The cores of $\tilde{\mu}, \tilde{\tau}_i, \tilde{a}_1, ..., \tilde{a}_m$ are estimated using the ordinary least squares (OLS) regression method. The estimated cores are denoted $\hat{\mu}, \hat{\tau}_i, \hat{a}_1, ..., \hat{a}_m$. Note that $\tilde{\mu} = \hat{\mu}$ and $\tilde{\tau}_i = \hat{\tau}_i$ for each i.

Step 2: Set $Y_j^* = Y_{ij} - \hat{\mu} - \hat{\tau}_i$. Note that index i is omitted as its effect has been corrected. The resulting regression equation is (24), which is rewritten in the standard way as:

$$\tilde{Y}_j^* = \tilde{a}_1 X_{1j} + ... + \tilde{a}_m X_{mj}$$

Step 3: The left and right spreads of $\tilde{a}_1, \tilde{a}_2, ..., \tilde{a}_m$ above are estimated from the same optimization as described in Section 2 without the intercept, for a given minimum accomplishment level α^*.

Step 4: Finally the fuzzy regression for the ith group or treatment is given by

$$\tilde{Y}_j = \hat{\mu} + \hat{\tau}_i + \tilde{a}_1 X_{1j} + ... + \tilde{a}_m X_{mj} \tag{26}$$

where index j runs over the observations in the ith group.

4.3. Finding past claim trend using fuzzy ANCOVA. We apply the general procedures for the FR developed in the previous subsection for the working data with the fuzzy regression equation of the log Hoerl curve

$$\tilde{Y}_{ij} = \tilde{c} + \tilde{\alpha}_i + \tilde{\beta} \log(j) + \tilde{\gamma} j \tag{27}$$

where $\tilde{Y}_{ij} = \log(C_{i,j})$. The result of Step 1 has already been done using the crisp OLS with the estimates given in (16). For Step 2, we set $Y_j^* = Y_{ij} - \hat{c} - \hat{\alpha}_i$ and

$$\tilde{Y}_j^* = \tilde{\beta} \log(j) + \tilde{\gamma} j \tag{28}$$

In Step 3, for the upper and lower limit of Y_j, we naturally set

$$\overline{Y_j} = \max_{\forall i} \{\log(C_{i,j}) - \hat{c} - \hat{\alpha}_i\}$$

and

$$\underline{Y_j} = \min_{\forall i} \{\log(C_{i,j}) - \hat{c} - \hat{\alpha}_i\},$$

where $i = 1, ..., 8$. At $\alpha^* = 0.3$, we obtain the two TFN parameters:

$$\tilde{\beta} = (\hat{\beta}, l_\beta, r_\beta) = (1.85624, 0.0000, 0.8978)$$

and

$$\tilde{\gamma} = (\hat{\gamma}, l_\gamma, r_\gamma) = (-1.31755, 0.3286, 0.0268).$$

From (18) the fuzzy projection rate is given by

$$\tilde{f}_{j,s} = \frac{\sum_{k=1}^{s} \exp(\tilde{\beta} \log k + \tilde{\gamma} k)}{\sum_{k=1}^{j} \exp(\tilde{\beta} \log k + \tilde{\gamma} k)} \tag{29}$$

We note that $f_{j,s}$ is a non-linear function of β and γ (the intercept term $c + \alpha_i$ disappears after canceling out in the ratio), and that the resulting FN $\tilde{f}_{j,s}$ is not a TFN due to nonlinearity, warranting a linear approximation. In order to use the first-order Taylor expansion, as described in Section 2, we obtain the partial derivatives of $f_{j,s}$ in (18) as

$$\frac{\partial f_{j,s}}{\partial \beta} = \frac{1}{[\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)]^2} \Big\{ [\sum_{k=1}^{s} \log k \exp(\beta \log k + \gamma k)] \cdot [\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)]$$

$$- \sum_{k=1}^{s} \exp(\beta \log k + \gamma k) \cdot [\sum_{k=1}^{j} \log k \exp(\beta \log k + \gamma k)] \Big\} \tag{30}$$

and

$$\frac{\partial f_{j,s}}{\partial \gamma} = \frac{1}{[\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)]^2} \Big\{ [\sum_{k=1}^{s} k \exp(\beta \log k + \gamma k)] \cdot [\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)]$$
$$- \sum_{k=1}^{s} \exp(\beta \log k + \gamma k) \cdot [\sum_{k=1}^{j} k \exp(\beta \log k + \gamma k)] \Big\} \tag{31}$$

In addition, one can also show that, for $j < s$, both of these partial derivatives are positive. To prove this, we denote $g(k) = \exp(\beta \log k + \gamma k)$, which is always positive, for notational convenience. Then, from (30),

$$\begin{aligned}
\frac{\partial f_{j,s}}{\partial \beta} &= \frac{1}{[\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)]^2} \Big\{ [\sum_{i=1}^{s} \log i \ \exp(\beta \log i + \gamma i)] \cdot [\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)] \\
&\quad - \sum_{i=1}^{s} \exp(\beta \log i + \gamma i) \cdot [\sum_{k=1}^{j} \log k \exp(\beta \log k + \gamma k)] \Big\} \\
&= \frac{1}{[\sum_{k=1}^{j} g(k)]^2} \Big\{ \sum_{i=1}^{s} \log i \, g(i) \cdot \sum_{k=1}^{j} g(k) - \sum_{i=1}^{s} g(i) \cdot \sum_{k=1}^{j} \log k \, g(k) \Big\} \\
&= \frac{1}{[\sum_{k=1}^{j} g(k)]^2} \sum_{i=1}^{s} \sum_{k=1}^{j} (\log i - \log k) \, g(i) \, g(k) \\
&= \frac{1}{[\sum_{k=1}^{j} g(k)]^2} \Big\{ \sum_{i=1}^{j} \sum_{k=1}^{j} (\log i - \log k) \, g(i) \, g(k) + \sum_{i=j+1}^{s} \sum_{k=1}^{j} (\log i - \log k) \, g(i) \, g(k) \Big\} \\
&= \frac{1}{[\sum_{k=1}^{j} g(k)]^2} \sum_{i=j+1}^{s} \sum_{k=1}^{j} (\log i - \log k) \, g(i) \, g(k) > 0
\end{aligned}$$

The last inequality holds because $\log i - \log k > 0$ for all values for $k = 1, ..., j$ and $i = j + 1, ..., s$.

Similarly, for (31),

$$\begin{aligned}
\frac{\partial f_{j,s}}{\partial \gamma} &= \frac{1}{[\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)]^2} \Big\{ [\sum_{i=1}^{s} i \ \exp(\beta \log i + \gamma i)] \cdot [\sum_{k=1}^{j} \exp(\beta \log k + \gamma k)] \\
&\quad - \sum_{i=1}^{s} \exp(\beta \log i + \gamma i) \cdot [\sum_{k=1}^{j} k \exp(\beta \log k + \gamma k)] \Big\} \\
&= \frac{1}{[\sum_{k=1}^{j} g(k)]^2} \Big\{ \sum_{i=1}^{s} i \, g(i) \cdot \sum_{k=1}^{j} g(k) - \sum_{i=1}^{s} g(i) \cdot \sum_{k=1}^{j} k \, g(k) \Big\} \\
&= \frac{1}{[\sum_{k=1}^{j} g(k)]^2} \sum_{i=1}^{s} \sum_{k=1}^{j} (i - k) \, g(i) \, g(k) \\
&= \frac{1}{[\sum_{k=1}^{j} g(k)]^2} \Big\{ \sum_{i=1}^{j} \sum_{k=1}^{j} (i - k) \, g(i) \, g(k) + \sum_{i=j+1}^{s} \sum_{k=1}^{j} (i - k) \, g(i) \, g(k) \Big\}
\end{aligned}$$

$$= \frac{1}{[\sum_{k=1}^{j} g(k)]^2} \sum_{i=j+1}^{s} \sum_{k=1}^{j} (i-k)\, g(i)\, g(k) > 0.$$

Now keeping in mind the signs of the partial derivatives, the approximated TFN, denoted $\tilde{f}'_{j,s} = (f_{j,s}, l_{f_{j,s}}, r_{f_{j,s}})$, is given by, from (8),

$$l_{f_{j,s}} = \frac{\partial f_{j,s}}{\partial \beta}\, l_\beta + \frac{\partial f_{j,s}}{\partial \gamma}\, l_\gamma; \qquad r_{f_{j,s}} = \frac{\partial f_{j,s}}{\partial \beta}\, r_\beta + \frac{\partial f_{j,s}}{\partial \gamma}\, r_\gamma \tag{32}$$

Using the estimated values for each j and s, one can readily calculate all the future projection rates that can help fill the run-off table.

4.4. Fuzzy claim reserve. Recall that in the crisp approach, the future cumulative loss $Z_{i,s}$, $s > n-i+1$, was given by

$$Z_{i,s} = Z_{i,n-i+1} \times f_{n-i+1,s} \tag{33}$$

In the fuzzy approach, therefore, we would use the linear approximation of the FN

$$\begin{aligned} \tilde{Z}_{i,s} &= Z_{i,n-i+1} \times \tilde{f}'_{n-i+1,s} = Z_{i,n-i+1} \times (f_{n-i+1,s}, l_{f_{n-i+1,s}}, r_{f_{n-i+1,s}}) \\ &= (Z_{i,n-i+1}\, f_{n-i+1,s}, Z_{i,n-i+1}\, l_{f_{n-i+1,s}}, Z_{i,n-i+1}\, r_{f_{n-i+1,s}}) \\ &= (Z_{i,s},\ \ l_{Z_{i,s}},\ \ r_{Z_{i,s}}), \end{aligned} \tag{34}$$

which is a TFN. Therefore the fuzzy reserve for accident year i is given by

$$\tilde{R}_i = \tilde{Z}_{i,n} - Z_{i,n-i+1} = (Z_{i,n} - Z_{i,n-i+1},\, l_{Z_{i,n}},\, r_{Z_{i,n}}) \tag{35}$$

and the total reserve by

$$\tilde{R} = (R, l_R, r_R) = \sum_{i=1}^{n} \tilde{R}_i = \left(\sum_{i=1}^{n} (Z_{i,n} - Z_{i,n-i+1}), \sum_{i=1}^{n} l_{Z_{i,n}}, \sum_{i=1}^{n} r_{Z_{i,n}} \right) \tag{36}$$

For our run-off cumulative claim data (2), we present the fuzzy values (34) of $\tilde{Z}_{i,s}$ in Table 3, and the reserves in Table 4. To look at the trend from the past we also included the past Z_{ij} values in the top panel of Table 3.

From Table 4 the total fuzzy reserve is given by

$$\tilde{R} = (R, l_R, r_R) = (659.79, 473.79, 618.5),$$

meaning that the claim reserve for this portfolio is approximately 659.79 but there may be deviation below (above) no greater than 473.79 (618.5). We can draw similar conclusion for other choices of α^* with smaller values leading to smaller spreads for both cumulative claims and the reserves.

5. Concluding remarks

In non-life insurance applications determining the evolution of the future claims is an important task to calculate the reserve which must be set aside to protect the insurer from future losses. The traditional claim reserving approach in the insurance literature typically uses a parametric regression model to estimte the future claim amounts and thus obtain the reserve. However this

TABLE 3. Predicted fuzzy value of cumulative loss Z_{ij}: Center, left and right spreads

		Development yr (j)							
	Accid yr (i)	1	2	3	4	5	6	7	8
Z_{ij}	1	101	254	306	323	337	340	344	345
	2	99	220	296	328	338	341	342	342.55
	3	110	292	372	392	413	415	416.95	417.61
	4	160	357	439	477	496	502.61	504.97	505.78
	5	161	415	500	546	566.09	573.64	576.33	577.25
	6	185	386	472	519.14	538.24	545.42	547.97	548.85
	7	178	439	561.84	617.95	640.69	649.23	652.27	653.32
	8	168	330.89	423.48	465.77	482.91	489.34	491.64	492.43
$l_{Z_{ij}}$	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	1.04
	3	0	0	0	0	0	0	3.09	4.36
	4	0	0	0	0	0	8.42	12.19	13.76
	5	0	0	0	0	19.67	29.54	33.95	35.76
	6	0	0	0	33.74	53.69	63.54	67.89	69.67
	7	0	0	60.86	107.1	133.3	145.96	151.47	153.7
	8	0	53.52	114.37	156.07	178.59	189.17	193.69	195.5
$r_{Z_{ij}}$	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0.79
	3	0	0	0	0	0	0	2.54	3.5
	4	0	0	0	0	0	7.62	10.73	11.92
	5	0	0	0	0	20.05	29.02	32.66	34.05
	6	0	0	0	39.97	60.51	69.59	73.25	74.63
	7	0	0	88.49	144.91	172.94	185.09	189.93	191.74
	8	0	105.73	202.02	258.05	284.65	295.87	300.25	301.87

TABLE 4. Fuzzy reserves for each accident year

accid yr (i)	R_i	l_{R_i}	r_{R_i}
1	0	0	0
2	0.55	1.04	0.79
3	2.61	4.36	3.5
4	9.78	13.76	11.92
5	31.25	35.76	34.05
6	76.85	69.67	74.63
7	214.32	153.7	191.74
8	324.43	195.5	301.87
sum	659.79	473.79	618.5

crisp approach is criticized on statistical bases including the violation of the error assumptions required for regression analyses, and the fuzzy regression method can serve as an alternative solution. In the present article we extend the fuzzy claim reserving method of [8] where the simple fuzzy regression is adopted ignoring the cohort effect in the claim data. We develop a fuzzy counter part of

the well-known Hoerl curve which accommodates the cohort effect as well as the development periods. This task, however, also calls for a fuzzy counterpart of the analysis of covariance, or ANCOVA, a combination of the linear regression and the analysis of variance. Our proposed fuzzy ANCOVA is simple to use and consistent with the statistical ANCOVA. Using an actual insurance claim data we find the fuzzy Hoerl curve adequately calculates relevant claim reserving quantities.

References

1. D. Dubois and H. Prade, *Analysis of Fuzz Information*, Vol. 2, chapter Fuzzy numbers: an overview. CRC-Press, Boca Raton (1988).
2. P. England and R. Verrall, Stochastic claims reserving in general insurance. *British Actuarial Journal*, Vol. 8 (2002), No. 3, 443–518.
3. H. Ishibuchi and M. Nii, Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks. *Fuzzy Sets and Systems*, Vol. 119 (2001), No. 2, 273–290.
4. R. Kaas, *Modern actuarial risk theory: Using R*, Springer, New York, 2008.
5. A. Kumar and A. Kaur, Application of linear programming for solving fuzzy transportation problems, *Journal of Applied Mathematics and Informatics*, Vol. 29 (2011), No. 3–4, 831–846.
6. H. Maleki and M. Mashinchi, Fuzzy number linear programming: A probabilistic approach (3), *Journal of Applied Mathematics and Informatics*, Vol. 15 (2004), No. 1–2, 333–341.
7. J. Neter, M. Kutner, C. Nachtsheim, and W. Wasserman, *Applied linear statistical models*. Irwin, Chicago, 4th edition, 1996.
8. J. Sánchez, Calculating insurance claim reserves with fuzzy regression. *Fuzzy sets and systems*, Vol. 157 (2006), No. 23, 3091–3108.
9. R. Sherman, Extrapolating, smoothing, and interpolating development factors. In *Proceedings of the Casualty Actuarial Society*, Vol. 71 (1984), 122–155.
10. H. Tanaka, S. Uejima, and K. Asai, Linear regression analysis with fuzzy model. *IEEE Trans. Systems Man Cybern*, Vol. 12 (1982), 903–907.
11. T. Wright, A stochastic method for claims reserving in general insurance. *Journal of the Institute of Actuaries* Vol. 117 (1990), 677–731.

Joseph H. T. Kim received M.S. and Ph.D. in Actuarial Science from University of Waterloo. He is currently an associate professor at Yonsei University since 2012. His research interests is quantitative risk management in insurance and finance.

Department of Applied Statistics and Quantitative Risk Management (QRM), College of Business and Economics, Yonsei University, Seoul 120-749, Korea.
e-mail: jhtkim@yonsei.ac.kr

Joocheol Kim received M.S from Korea Advanced Institute of Science and Technology, and Ph.D. from Georgia Institute of Technology. He is currently an associate professor at Yonsei University since 2003. His research interests are financial engineering, risk management and fuzzy regression.

Department of Economics and Quantitative Risk Management (QRM), College of Business and Economics, Yonsei University, Seoul 120-749, Korea.
e-mail: joocheol@yonsei.ac.kr