Prediction of Loss Ratio Using Nonparametric Regression

Prepared by

Dr. Mahmoud Mohamed Elsayed
Associate professor
Insurance Department
Faculty of Commerce – Cairo University

Dr. Amr Soliman Mohamed
Lecturer of Insurance
Insurance Department
Faculty of Commerce – Cairo University

Scientific Journal for Financial and Commercial Studies and Researches (SJFCSR)

Faculty of Commerce – Damietta University

Vol.2, No.1, Part 1., Jan. 2021

APA Citation:

EL Sayed, M. M. and **Mohamed**, A. S. (2021). Prediction of Loss Ratio Using Nonparametric Regression**.** ***Scientific Journal for Financial and Commercial Studies and Researches*** *(SJFCSR)*, Vol.2 (1) Part1. pp.527- 549.

Website: **https://cfdj.journals.ekb.eg/**

Abstract

The aim of this paper is to use the nonparametric regression to predict the loss ratio of non-life insurance as a dependent variable, based on incurred claims and earned premiums as two explanatory variables. Secondary data adopted in this research consists of twelve year time series of incurred claims and earned premiums from 2007/2008 to 2018/2019. In this paper we used R Package to apply kernel smoothing and estimation of the smoothing parameter (bandwidth) using cross validation methods. Afterwards, we applied a local polynomial (constant and linear) and Locally Estimated Smoothing (LOESS) in order to predict loss ratio. The results from local polynomial and LOESS indicate that the predicted value of loss ratio has a slight increasing trend over the upcoming three years period. All analysis and calculations in this paper has been performed using the R package codes as open source software.

Keywords: Nonparametric Regression, Kernel Smoothing, Bandwidth, Insurance, Loss Ratio. **JEL Classification:** C14, C53, G22.

1. Introduction

Insurance market is one of the most vital financial institutions that contribute to the Egyptian GDP, insurance sector is considered as a tool for the stability and continuity of the Egyptian economy. In 2019, there are 39 insurers operating in the Egyptian insurance industry divided into 24 nonlife insurers and 15 life insurers.

According to the Egyptian Financial Regulatory Authority (FRA) reports, the insurance industry represents 0.90% of the Egyptian GDP in 2019 compared to 0.91% in 2018. The growth rate of insurance premiums increased by 19.7% in 2019 compared to 2018. The total amount of premiums increased from 35.2 billion EGP in 2019 where the total premiums in 2018 were 29.4 billion EGP. Whereas, total amount of claims paid in 2019 were 18.3 billion EGP compared to 15.4 billion EGP in 2018 with a growth rate of 18.9%. On the other hand, the surplus increased by 29.3% in 2019 compared to 2018 that leads to attracting new investment in the insurance industry. In the same vein, the net investment increased from 99.4 billion EGP in 2018 to 102 billion EGP in 2019 with a growth rate of 2.6%. Moreover, the total rights of policyholder’s amount to 69 billion EGP on 2019 compared to 61 billion EGP in 2018 with a growth rate of 13%. While the shareholders equity amount to 35 billion EGP in 2019 compared to 38 billion EGP in 2018 with a declined rate of 7.8%.

Nonparametric approach is a statistical technique that doesn't require any predetermined form or relation between variables, while the data itself forms the resulting model. Moreover, nonparametric techniques give more flexibility to deal with data than parametric techniques, nonparametric techniques also treat the data without making any assumptions compared to parametric methods. Furthermore, nonparametric methods can be combined with parametric models to establish semi-parametric models. Recently, nonparametric methods becoming increasingly popular in applied statistical methods and data analysis. Nonparametric methods can be used in regression analysis and this approach is called nonparametric smoothing. Nonparametric regression can be used for different tasks such as prediction, smoothing, fitting and testing hypothesis. In this paper we used kernel smoothing and local polynomial to estimate the regression function as nonparametric regression estimators. In addition, we measured the performance of the estimators by Mean Integrated Squared Error (MISE) and Asymptotic Mean Integrated Squared Error (AMISE).

The construction of the research will be as follows: next section will be literature review, third section will be data and methodology, forth section related to the models, fifth section discussed the empirical results and finally the conclusion.

2. Literature Review

Lopes, et al (2012) introduced a nonparametric technique for creating an additional reserve so called Incurred but Not Reported (IBNR) based on kernel smoothing, support vector regression and Gaussian process regression. Furthermore, they compared these results to benchmark IBNR estimation model and Mack's chain ladder and concluded that proposed models produce closest prediction of IBNR. Moreover, Delaigle and Gijbels (2004b) applied kernel estimation and selection of bandwidth using several methods: plug-in method, cross validation method and bootstrap method. They concluded that plug-in and bootstrapping methods perform comparably and both outperform the cross validation method. In addition, Gong and Gao (2017) proposed a bootstrapping technique to estimate the confidence bands for discontinuity and evaluated the estimated results by Monte Carlo simulation. This paper applied nonparametric kernel estimation on Medicare levy Surcharge in Australia.

Ceglia (2016) used parametric and nonparametric methods to select the variables that affect the fraudulent claims by logistic regression, Least Absolute Shrinkage and Selection Operator method (LASSO) and random forests. This paper concluded that random forest yields highest accuracy to predict the fraudulent claims. In the same vein, Taha, et al (2011) applied parametric methods to predict loss reserve for general insurance market, they used three methods for prediction and these methods are: exponential smoothing, Box-Jenkins and time series regression. They concluded that exponential smoothing is defined as the best forecasting technique. However, Delaigle and Meister (2007) proposed a nonparametric regression when heteroscedastic errors in variable are identically distributed; they also give an optimal regression function. Furthermore, Gijbels and Goderniaux (2004) introduced a bootstrap method for selecting smoothing parameters for nonparametric regression. This paper succeeded to reach a fully data-driven method to estimate a finite number of change points in regression.

Jeffrey (2008) presents semi-parametric and nonparametric techniques and demonstrates several methods used for smoothing and bandwidth selection. Also, this paper emphasized the computational methods using R package. In addition, Delaigle and Gijbels (2004a) and (2004b) proposed a bootstrap technique for optimal bandwidth selection for ordinary kernel density estimation. They applied a simulation study to test the performance of the selection.

In this paper we are interested in prediction of nonlife insurance loss ratio based on nonparametric regression as a dependent variable, while incurred claims and earned premiums as explanatory variables.

3. Data and Methodology

Nonparametric regression is applied to predict loss ratio as a dependent variable depending on both incurred claims and earned premiums as explanatory variables. Data adopted in this research reported in Financial Regulatory Authority (FRA) for nonlife Egyptian insurance market.

Secondary data consists of twelve year time series from 2007/2008 to 2018/2019 of loss ratio, incurred claims and earned premiums.

Table (1) describes the trend of incurred claims and earned premiums for the Egyptian nonlife insurance market during the period from 2007/2008 to 2018/2019. The table also presents the loss ratio (Incurred Claims divided by Earned premiums) for the same period.

Table 1 Total amount of incurred claims, earned premiums for nonlife insurance Egyptian market and the market loss ratio.

In Thousands (EGP)

Year	Incurred Claims	Earned Premiums	Loss Ratio
2007/2008	2409495	1897567	1.27
2008/2009	2080829	2491213	0.84
2009/2010	2728816	2884237	0.95
2010/2011	2652673	3217863	0.82
2011/2012	2217705	3168890	0.70
2012/2013	2470999	3434897	0.72
2013/2014	2215532	3908968	0.57
2014/2015	2747947	4637290	0.59
2015/2016	2561639	5397626	0.47
2016/2017	3312489	7218110	0.46
2017/2018	5119990	8849149	0.58
2018/2019	6365221	10975058	0.58

Source: Financial Regulatory Authority (FRA).

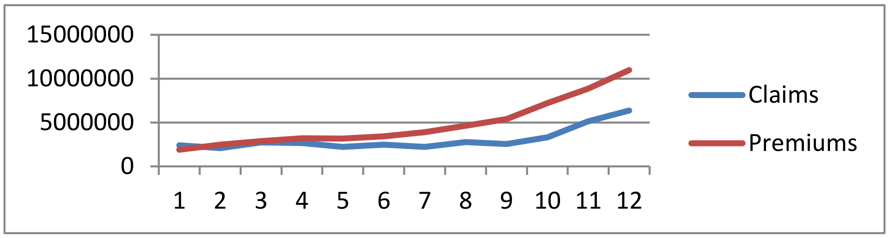

Figure 1 Total amount of incurred claims and earned premiums for nonlife insurance Egyptian market.

From table (1) and figure (1) we can conclude that there is an increasing trend in premiums, but there is a drop in 2011/2012 according to the unstable economic situations because of the 2011 revolution.

In this paper we used the R Package to apply Gaussian kernel smoothing. Moreover, we estimated the bandwidth of the model using least squares cross-validation and likelihood cross-validation methods. Afterwards, we used a local polynomial (constant and linear) and loess in order to predict loss ratio.

4. Models

4.1 Nonparametric Regression

Nonparametric regression is a methodology for describing the trend between a response variable and one or more predictors. The main goal of nonparametric regression is to construct a mathematical model in order to estimate $\hat{y}$ from $i.i.d$ samples (x_i, y_i) for $i = 1,2,\dots,n$ where $(X,Y) \in R^d \times R$. Taking into consideration that no predetermined form or relation between x and y.

According to Wasserman (2006) the nonparametric regression model can be expressed as:

$$\hat{y}_i = m(x_i) + \varepsilon_i$$

$$m(x) = E[Y|X = x]$$

$$m(x) = \int \frac{y.f_{x,y}(x,y)}{f(x)}\, dy$$

where $\varepsilon_i \sim N(0, \sigma^2)$, ε_i is the white noise with zero mean and finite variance. The function $m(x)$ is the nonparametric regression function, also $m(x)$ doesn't make any parametric assumptions.

4.2 Kernel Smoothing

For simplicity let's assume a one dimensional model ($d = 1$) the kernel regression or kernel smoothing $K(x) > 0$, $K : R \to R$ satisfies Tibshirani and Wasserman (2015):

$$K(-x) = K(x) \;, \quad \int K(x)\, dx = 1 \;, \quad \int x\, K(x)\, dx = 0 \;,$$

$$0 < \int x^2\, K(x)\, dx < \infty$$

There are three commonly used models for smoothing nonparametric regression.

(1) Gaussian kernel: $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$

(2) Box kernel: $K(x) = \begin{Bmatrix} 1/2 & if\ |x| \le 1 \\ 0 & Otherwise \end{Bmatrix}$

(3) Epanechnikov kernel: $K(x) = \begin{Bmatrix} 3/4(1-x^2) & if\ |x| \le 1 \\ 0 & Otherwise \end{Bmatrix}$

Moreover, Nadaraya-Watson estimator is defined by

$$\widehat{m}(x) \equiv \widehat{m}_h(x) = \frac{\sum_{i=1}^{n} y_i\, K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

where $h > 0$ is the bandwidth of the estimate.

In a bivariate case or a 2 dimensional model $(d = 2)$ data points can be represented by two vectors $x_1 = [x_{11}, x_{12}, \ldots, x_{1n}]$ and $x_2 = [x_{21}, x_{22}, \ldots, x_{2n}]$ and $x_i = (x_{1i}, x_{2i})$. The bivariate kernel density function is defined by Bilock et al (2016)

$$\hat{f}(x, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(x - x_i)$$

Where **H** is the bandwidth positive definite matrix

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$

And $K_{\mathbf{H}}$ is a bivariate kernel density function having the same characteristics as $K(x)$.

4.3 Bandwidth Selection

In practice, there are several methods to select the bandwidth for smoothing the kernel function, these methods are:

1. Reference rules of thumb
2. Plug-in methods
3. Cross Validation methods
4. Bootstrap methods

The change in bandwidth has more effect on the fitness of the model than kernel function. In the same vein, the larger the value of h leads to more smoothed fits and vice versa.

4.3.1 Reference Rule-of-Thumb

Reference rule of thumb is used via a standard family distribution and it is commonly used for univariate density functions.

4.3.2 Plug-in methods

Plug-in methods also used for univariate density functions, it reflects uncertainty and insufficient manner that results inefficient estimates.

4.3.3 Cross-Validation (CV)

4.3.3.1 Least squares cross-validation

Least squares cross-validation method estimates the bandwidth that minimizes he integrated mean square error (IMSE), it is also considered as a fully automatic and data driven method. The integrated mean square error between $\hat{f}(x)$ and $f(x)$ see Jeffrey (2008)

$$\int \{\hat{f}(x) - f(x)\}^2 \, dx$$

4.3.3.2 Likelihood cross-validation

Likelihood cross-validation method is also called Kullback–Leibler estimate. It estimates h by maximize the log likelihood function as in Jeffrey (2008)

$$\mathcal{L} = \log L = \sum_{i=1}^{n} \log \hat{f}_{-i}(x)$$

where $\hat{f}_{-i}(x)$ is the leave-one-out kernel estimator of $f(X_i)$ that uses all points except X_i to construct the density estimate, that is

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h} \sum_{j=1\,, j\neq i}^{n} K\left(\frac{X_j - x}{h}\right)$$

4.3.4 Bootstrap Methods

Bootstrap approach based on a probabilistic or stochastic density function. Moreover, it minimizes the average square error of kernel density function in order to estimate the bandwidth.

4.4 Local polynomial Regression

Local polynomial Regression is used to evaluate the regression function at certain x - value x_0. Then fit the model based on different x- values over n observations.

The regression function can be written as a p^{th} – order weighted least square polynomial of y on x. Furthermore, the larger value of p gives more flexible smoothness Fox (2002).

$$y_i = \beta_0 + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 + \beta_3(x_i - x_0)^3 \ldots + \beta_p(x_i - x_0)^p + \varepsilon_i$$

We adopted $W(z)$ as a tricube function for weighting the observation to a focal value x_0.

where

$$W(z) = \begin{Bmatrix} (1 - |z|^3)^3 & for\ |z| < 1 \\ 0 & for\ |z| \geq 1 \end{Bmatrix}$$

And $z_i = (x_i - x_0)/h$

4.5 LOESS/LOWESS

LOESS (Locally estimated scatterplot smoothing) or LOWESS (Locally weighted scatterplot smoothing). Loess is a non-parametric approach that fits bivariate or multivariate regressions in local neighborhood. Loess regression can be applied to predict values of y based on numerical vectors $x_1 = [x_{11}, x_{12}, \ldots, x_{1n}]$ and $x_2 = [x_{21}, x_{22}, \ldots, x_{2n}]$ .The size of the neighborhood can be controlled using the span $s \in [0,1]$ argument. It controls the degree of smoothing. Therefore, the greater the values of s, the smoother is the curve.

5. Empirical Study

5.1 Descriptive Statistics

We will start this section by summarize the statistical characteristics of each variable (incurred claims, earned premiums as explanatory variables and loss ratio as dependent variable) as shown in table (2).

Table 2 Descriptive analysis of incurred claims, earned premiums and loss ratio

Measure	Incurred Claims	Earned Premiums	Loss Ratio
Mean	3073611.25	4840072.33	0.71
Standard Deviation	1314338.86	2795848.40	0.23
Minimum	2080829	1897567	0.46
Q_1	2361547.5	3097726.75	0.58
Median	2080829	1897567	0.46
Q_3	2889082.5	5852747	0.83
Maximum	6365221	10975058	1.27
Range	4284392	9077491	0.81
Skewness	1.98	1.25	1.34
Kurtosis	3.22	0.73	2.02
Standard Error	379417	807091.9	0.07

Source: Authors' Calculations.

5.2 Bandwidth Selection

In order to apply kernel density function, it is necessary to select the bandwidth of the smoothing, this bandwidth is also called the smoothing parameter **H** in the bivariate case. In this paper we adopt the maximum likelihood CV method and least square CV method.

The CV method can be applied to any kernel density function, in our research we implemented the R package code npcdensbw and the second order Gaussian function for bandwidth selection.

Table 3 Bandwidth selection based on likelihood CV and least square CV.

Bandwidth Selection Method	Maximum Likelihood Cross-Validation	Least Square Cross-Validation
Incurred Claims	124244.5	581667.3
Earned Premium	367812.1	1.232283e+13
Loss Ratio	0.1501515	0.0024366

Source: Authors' calculations based on results from R Package.

We implemented the R package codes to estimate kernel density function and selecting bandwidth parameters using the likelihood CV and least square CV as presented in table (3).

5.3 Goodness of fit of nonparametric regression

The goodness of fit can be assessed by calculating $\mathbf{R^2}$ which differs from the coefficient of determination in parametric models.

In nonparametric regression $\mathbf{R^2}$ is defined as:

$$\mathbf{R^2} = \frac{[\,\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})\,]^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$$

And $\mathbf{R^2}$ should lie between [0,1], when $\mathbf{R^2} = 1$ means there is a perfect fitting to the sample data, but when $\mathbf{R^2} = 0$ means no predictive power above that given by the unconditional mean of the outcome (Hayfield and Racine 2008). We applied the R Package codes npreg and loess to fit the loss ratio using local polynomial and LOESS with a span 0.99 as shown in table (4).

Table 4 fitting loss ratio using local polynomial, LOESS

Year	Actual Values	Fitted Values of Loss Ratio	
		Local Polynomial	**LOESS**
2007/2008	1.27	1.2211140	1.2692163
2008/2009	0.84	0.8394240	0.8352359
2009/2010	0.95	0.8804881	0.9516842
2010/2011	0.82	0.8291556	0.8257191
2011/2012	0.70	0.7136617	0.7007611
2012/2013	0.72	0.7452441	0.7182753
2013/2014	0.57	0.6163442	0.5654105
2014/2015	0.59	0.5802534	0.5455857
2015/2016	0.47	0.4890516	0.4366998
2016/2017	0.46	0.4589136	0.3994512
2017/2018	0.58	0.5785856	0.6214221
2018/2019	0.58	0.5799715	0.5801104

Source: Authors' calculations.

Afterwards, we implemented the R package codes summary, npreg and summary, loess to test the goodness of fit by $\mathbf{R^2}$ applied on kernel density function, and local polynomial. A second order Gaussian kernel is used and a local linear polynomial. The results obtained are shown in table (5).

Table 5 Goodness of fit of nonparametric regression

Measure	Local polynomial	LOESS
Residual Standard Error	0.006732723	0.0958
$\mathbf{R^2}$	0.9991199	0.99

Source: Authors' calculations.

5.4 Predicting loss ratio using nonparametric kernel regression

In this section we applied nonparametric kernel regression after selecting the optimal bandwidth by CV methods, in order to predict the loss ratio for nonlife Egyptian insurance market based on incurred claims and earned premiums. We performed a three years extrapolation for loss ratio as presented in tables (6) and (7).

Table 6 Predicted loss ratio using LOESS

Year	Point Estimate of Loss Ratio	95% Confidence Interval Predicted Loss Ratio	
		Lower Confidence Level	Upper Confidence Level
2019/2020	0.5227897	0.20543688	0.8401425
2020/2021	0.5748513	0.21916640	0.9305362
2021/2022	0.6269129	0.19730919	1.0565166

Source: Authors' calculations.

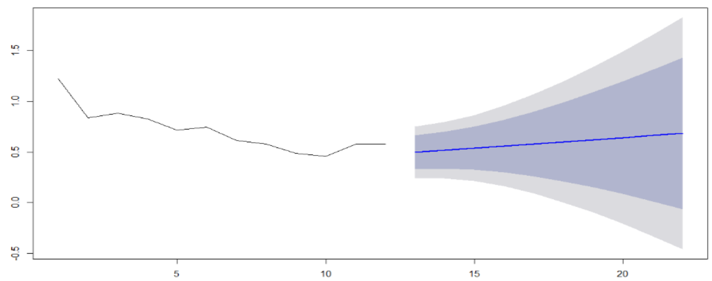

Figure 2 Predicted loss ratio using LOESS

Table 7 Predicted loss ratio using Local Polynomial

Year	Point Estimate of Loss Ratio	95% Confidence Interval Predicted Loss Ratio	
		Lower Confidence Level	Upper Confidence Level
2019/2020	0.4974655	0.24457326	0.7503578
2020/2021	0.5181087	0.24072002	0.7954973
2021/2022	0.5387518	0.21291155	0.8645920

Source: Authors' calculations.

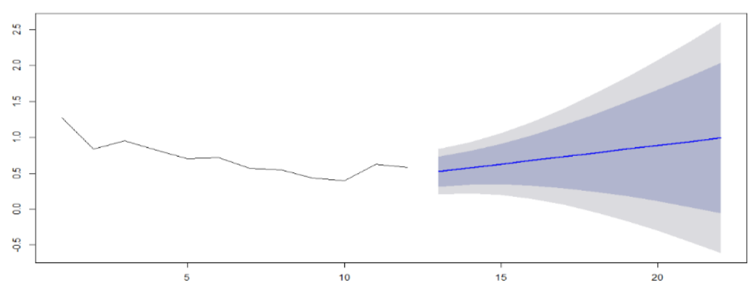

Figure 3 Predicted loss ratio using local polynomial

From tables (6), (7) and figures (2), (3) present the 95% confidence interval of the predicted loss ratio using LOESS and local polynomial.

6. Conclusion

Nonparametric regression is applied to predict loss ratio as a dependent variable depending on both incurred claims and earned premiums as explanatory variables. Data adopted in this research reported in Financial Regulatory Authority (FRA) for nonlife Egyptian insurance market. Moreover, we implemented the R package codes to estimate kernel density function and selecting bandwidth parameters using the likelihood CV and least square CV. Afterwards, we used a local polynomial and LOESS in order to predict loss ratio for the next three years.

The results from local polynomial indicate that the predicted value of loss ratio has a slight increasing trend over the three years period (52%, 57% and 63%) respectively. In addition, the results from LOESS show an increasing trend of the loss ratio (50%, 52% and 54%) respectively. However, we can conclude that both techniques move in the same direction in predicting the loss ratio.

References

Bilock, A. Jidling, C. and Rydin, Y (2016). Modelling Bivariate Distributions Using Kernel Density Estimation. Project in Computational Science. Department of information technology. UPPSALA UNIVERSITET.

Ceglia,C. (2016). A comparison of parametric and non-parametric methods for detecting fraudulent automobile insurance claims. Department of Mathematics and Statistics California State University, Long Beach. Msc thesis.

Delaigle, A. and Gijbels, I. (2004a). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. Annals of the Institute of Statistical Mathematics, 56(1): 19–47.

Delaigle, A. and Gijbels, I. (2004b). Practical bandwidth selection in deconvolution kernel density estimation. Computational statistics & data analysis, 45(2):249– 267.

Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association*, 102(480):1416–1426.

Fox, J. (2002). Nonparametric Regression. Appendix to An R and S-PLUS Companion to Applied Regression, Thousand Oaks CA: Sage.

Gijbels, I. and Goderniaux, A.-C. (2004). Bandwidth selection for changepoint estimation in nonparametric regression. Technometrics, 46(1):76–86.

Gong , X.and Gao,J.(2017). Nonparametric Kernel Estimation of the Impact of Tax Policy on the Demand for Private Health Insurance in Australia.

Working Paper. Department of Econometrics and Business Statistics, Monash University.

Hayfield T, Racine JS (2008). Nonparametric Econometrics: The np Package. *Journal of Statistical Software*, 27(5): 1-32.

Jeffrey.S. (2008). Nonparametric Econometrics: A Primer. Foundations and Trends in Econometrics, 3(1): 1–88.

Lopes, H. Barcellos, J. Kubrusly, J. and Fernandes, C. (2012). A non-parametric method for incurred but not reported claim reserve estimation. *International Journal for Uncertainty Quantification*, 2 (1): 39–51.

Taha, T. Ibrahim. Y. and Minai. M. (2011). Forecasting general insurance loss reserves in Egypt. *African Journal of Business Management*, 5(22): 8961-8970.

Tibshirani, R. and L. Wasserman (2015). Nonparametric Regression Statistical Machine Learning.

Wasserman, L. (2006). All of Nonparametric Statistics, Springer, New York.

http://www.fra.gov.eg/content/efsa_en/eisa_pages_en/main_eisa_page_en.htm

التنبؤ بمعدل الخسارة باستخدام الانحدار اللامعلمي

د. محمود محمد السيد أحمد
أستاذ مساعد بقسم التأمين والعلوم الاكتوارية
كلية التجارة – جامعة القاهرة
mah.elsayed_njk@foc.cu.edu.eg

د. عمرو سليمان محمد
مدرس بقسم التأمين والعلوم الاكتوارية
كلية التجارة – جامعة القاهرة
s.amr@foc.cu.edu.eg

ملخص البحث

هدف البحث الى التنبؤ بمعدل الخسارة كمتغير مستقل في شركات تأمينات الممتلكات والمسئوليات في سوق التأمين المصري بالاعتماد على الأقساط المكتسبة والتعويضات التحميلية كمتغيرات تابعة، وذلك خلال الفترة من ٢٠٠٨/٢٠٠٧ الى ٢٠١٩/٢٠١٨ وذلك باستخدام برنامج R . وتطبيق البرنامج للتنبؤ بتمهيد كرنال وتقدير معلمات التمهيد باستخدام Cross validation methods وتطبيق a local polynomial (constant and linear) and Locally Estimated Smoothing (LOESS)

وتوصل البحث إلى زيادة معدل الخسارة بمقدار ضئيل خلال الثلاث سنوات القادمة لشركات تأمينات الممتلكات والمسئوليات في السوق المصري.

الكلمات المفتاحية: الانحدار اللامعلمي، تمهيد كرنال، معامل التمهيد، معدل الخسارة