Smolarek Leszek
Maritime University, Gdynia, Poland

Wachnicka Joanna
University of Technology, Gdańsk, Poland

Fuzzy regression approach to road safety analysis at regional level

Keywords

road safety, fatality rate, regions, factors, modelling, Europe, fuzzy function, Poisson model

Abstract

Road safety modelling on regional level of NUTS 2 in the EU is the complex issue ant authors of this article indicate this in previous publications. During multivariate models development they discovered that it is difficult to make regression model well described all regions, even if they are from one country. In the first step Poisson model of road fatality rate as fatalities per 100 thou of citizens was prepared. In this article was presented the fuzzy regression approach to estimate lover and upper bounds of the modeled value.

1. Introduction

The issue of road traffic safety has been studied on international level for many years now and many publications have appeared on the issue. It is a problem, which embraces many scientific fields seemingly not related to each other, such as road engineering, economy, mathematics, transport or medicine. Studies of changes occurring in traffic safety in respective countries were often based on analyses of changes in road fatality rate in a given time horizon [3], [7]. Sometimes researchers analyzed, which factors may influence observed changes [8]. Such kind of approach could be a source of information for legislative authorities or road administrators to be used for effective road traffic safety management.

However, as experience shows, the actions run at the national level translate into effects in lower administration units to variable extents. There might be an improvement of traffic safety observed in one region and none in the other. Therefore it is justified to analyze national as well as regional characteristics of a given area and their potential impact on the road safety level.

In the researches made at the national level gross national product per capita [10] and transport activity [9] are often listed as significant factors influencing traffic safety.

Authors in previous articles regression models for all European regions [11]. Then they made multilevel models including regional and national characteristics [12]-[13]. They assumed that modelled variable has Poisson distribution. In both approaches deterministic or stochastic models embraced part of data whereas rest of data lie beyond function line. To improve probability of good road safety estimation it was fuzzy regression made.

2. Methodology

Analyses of road safety at levels below the national level have usually compared selected rates of road safety (such as the road fatality rate in relation to demography). Fatality is calculated as follows:

$$FATALR_{ij} = \frac{FATAL_{ij}}{POP_{ij}} \tag{1}$$

where:
$FATALR_{ij}$ – traffic fatality rate in an *i*-th region in *j*-th year
$FATAL_{ij}$ – number of fatalities of an *i*-th region in j-th year
POP_{ij} – number of population of an *i*-th region in *j*-th year.
The same approach was used in this article.

FATALR estimation is based on the assumption that the parameter has Poisson distribution.
Fuzzy regression approach was made for road safety modelling on regional levels on Swedish data set.
Histograms of FATALR values in the regions of the compared countries within analysed period of 1999-2008 presented in *Figure 1* indicate that there is no ground to dismiss the hypothesis of Poisson distribution, which is frequently adopted in the analyses of safety level [6].

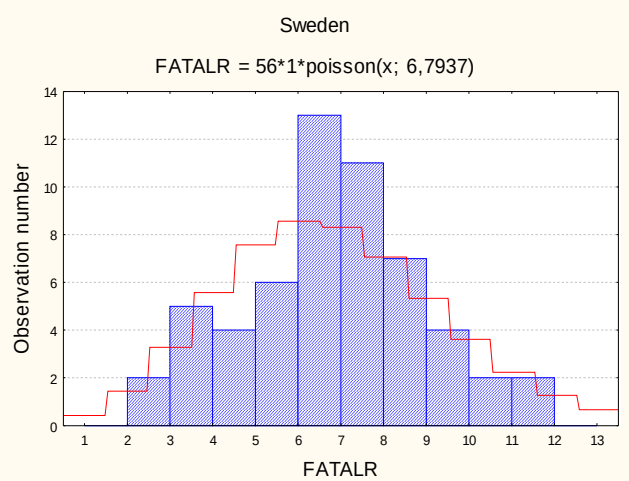

Figure 1. Histograms of analysed FATALR rates in regions of Sweden

In the article from 2012 year was developed mathematical model of road fatality rate FATALR consists of two variables: national product per capita NPPC and road concentration ROADC. In *Figure 2* was presented graph of real FATALR noted for Swedish regions and developed model. Model function lines reflected road concentration value as minimum, average and maximum from the set.

$$FATALR = 937{,}868 \cdot (\ln NPPC)^{-4{,}291} \cdot e^{(0{,}142 \cdot ROADC)} \qquad (2)$$

where:
FATALR – the road fatality rate in relation to demography in a given region [fatalities/100 thou. inhabitants]
ln*NPPC* – logarithm of national product per capita [-]
ROADC - road concentration [km/km^2]

As we can see in *Figure 2* model line doesn't embrace all cases, especially in the beginning of data set. It is essential to notice that national product per capita NPPC is the variable which diversifies strongly FATALR function. That's why without making big mistake this variable can be replaced by average from data set.
Next authors tried to find model for wider extend of data. For this purpose were made models which describes upper and lower edges of real FATALR data. These models don't contain ROADC variables because little impact of this variable was testified.

$$FATALR_{MAX} = NPPC^{1{,}401} \cdot e^{(-0{,}089 \cdot NPPC)} \qquad (3)$$

$$FATALR_{MIN} = 18{,}458 \cdot e^{(-0{,}047\ NPPC)} \qquad (4)$$

In *Figure 3* was presented two-factor model and models of data range bounds against the background of real data.

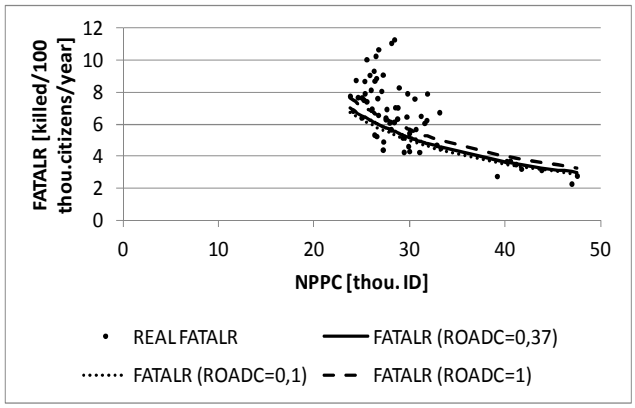

Figure 2. Graph of prepared regional FATALR model in relation to national product per capita and against the actual data - road concentration ROADC from the model assumed as minimum, average and maximum value from data set

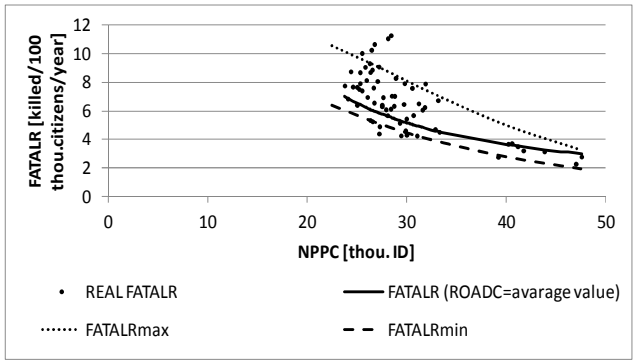

Figure 3. Graph of prepared regional FATALR model and bound models in relation to national product per capita and against the actual data

2. Fuzzy nonlinear regression with non-fuzzy input and fuzzy output data.

A regression analysis is a statistical technique applied to data to determine the best mathematical expression describing the functional relationship

between the response and the independent variables.
In traditional statistical analysis mode, the explanatory variables or the response variable is usually assumed to be a clear value, but in reality many observations often do not comply with this feature, since many of the variables cannot be explained by the exact value.
Statistical regression has many applications, but it is problematic if the data set is too small, or if there is vagueness in the relationship between the independent and dependent variables, or if the linearity assumption is inappropriate. These are the situations fuzzy regression was meant to address, [1].
Zadeh, [15] proposed the concept of fuzzy sets , the uncertainty of human cognitive process variables (mainly thinking and reasoning) as may be fuzzy numbers (Fuzzy Number) said , so in order to solve this problem.
Tanaka, [14], proposed fuzzy linear regression model by combining fuzzy theory and linear regression model.
Diamond, [4], regression model identified in the explanatory variables, the response variable and the error term regarded as fuzzy numbers , and then use the minimum number of α-cut fuzzy definition of its square criterion function to deal with the problem of information ambiguity.
The fuzzy regression model may be roughly classified by conditions of independent and dependent variables into three categories, as follows, [16]-[17]:

- Input and output data are both non-fuzzy number.
- Input data is non-fuzzy number but output data is fuzzy number.
- Input and output data are both fuzzy number.

In this paper, we present a new technique to obtain a fuzzy regression model, where the independent Variable (input data) NPPC is non fuzzy number but the response variable (output data) FATALR is fuzzy number. In the model we consider triangular fuzzy numbers in the response variable.

Definition 1. A triangular fuzzy number (TFN) *A* is a fuzzy number with a piecewise linear membership function f_A defined by:

$$f_A = \begin{cases} 0, x \le a \\ (x-a)/(b-a), x \in (a,b] \\ (c-x)/(c-b), x \in (b,c) \\ 0, x \ge c \end{cases} \quad (5)$$

It has three parameters '*a*' (minimum), '*b*' (middle) and '*c*' (maximum) that determine the shape of the triangle.

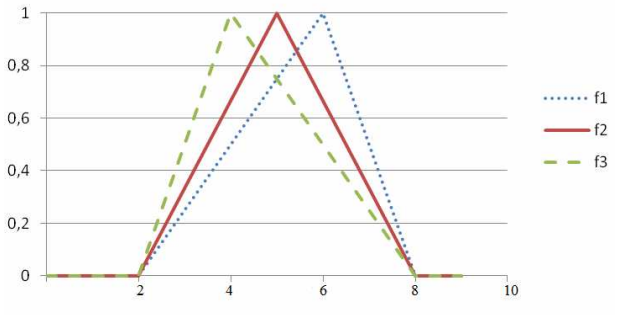

Figure 4. Examples of triangular membership functions(FATALR), where a=2, c=8 and b=4,5,6; NPPC=28-30

Form of a membership function:

In the article we introduce the concept of using the envelop functions to built TFN.

The methodology for our simple example proceeds in three steps:

- In the first step, we estimate the parameters of the classical multidimensional regression model using the dataset. This gives us regression function described by formula (2).
- In the second step, we estimate the both envelope functions. Using order statistics and regression analysis we obtain two functions depended on the variable NPPC, formula (3), (4).
- In the third step, we compute the parameters of a piecewise linear membership function for each NPPC value.

The membership function of a fuzzy set represents the degree of truth as an extension of valuation.
Degrees of truth are distinct according to probabilities because fuzzy truth represents only membership in vaguely defined sets.
The fuzzy response variable *FATALR* is a triangular fuzzy number, formula (5), for each NPPC value, *Figure 4*, with a piecewise linear membership function $f_{FATALR}(NPPC)$ defined by:

$$f_{FATALR}(NPPC) = \begin{cases} 0 & NPPC \le FATALR_{MIN} \\ \dfrac{NPPC - FATALR_{MIN}}{FATALR - FATALR_{MIN}} & FATALR_{MIN} < NPPC \le FATALR \\ \dfrac{FATALR_{MAX} - NPPC}{FATALR_{MAX} - FATALR} & FATALR < NPPC \le FATALR_{MAX} \\ 0 & FATALR_{MAX} < NPPC \end{cases} \quad (6)$$

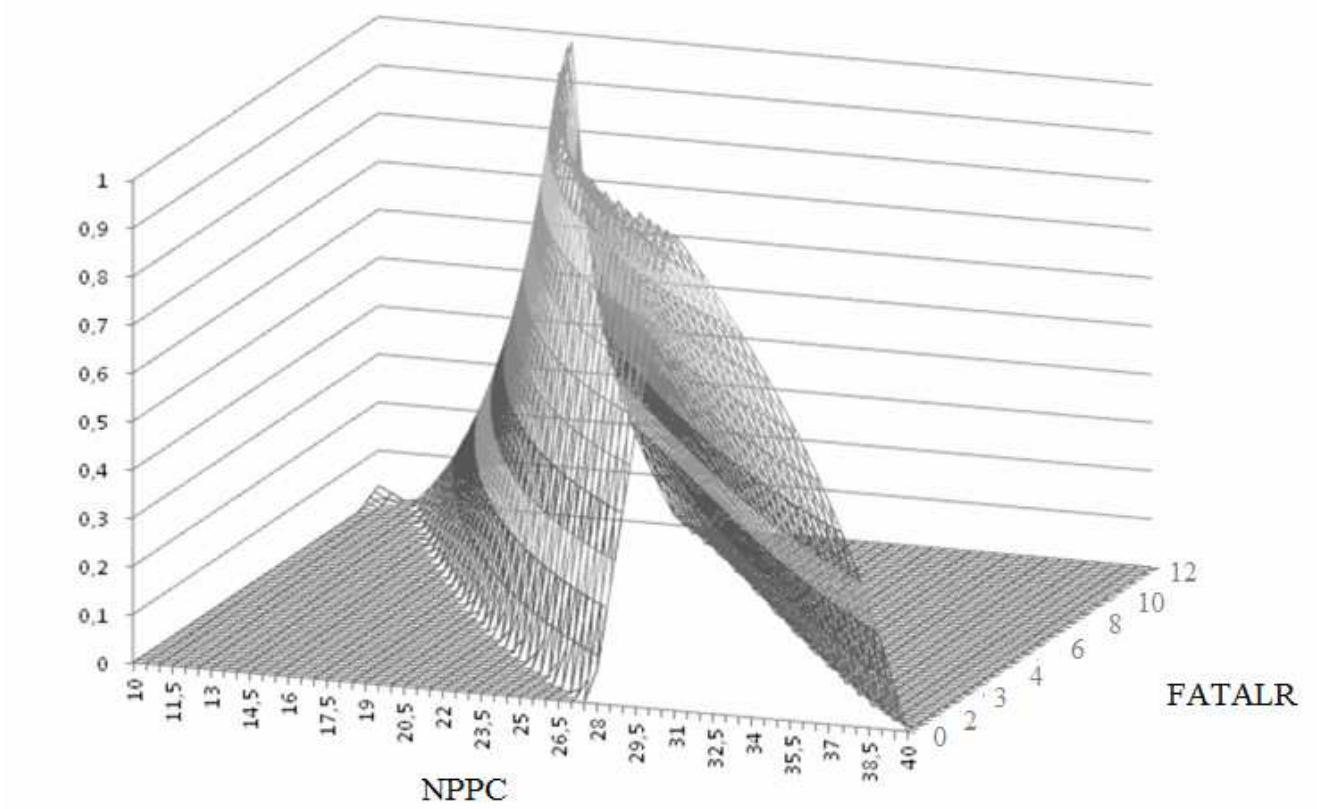

Figure 5 The fuzzy regression model with the triangular membership function f_{FATALR}(NPPC), formula (6)

Generally a membership function for a fuzzy interval A on the real axis X is defined as $f_A: X \rightarrow <0,1>$,

where each element of X is mapped to a value between 0 and 1, [1], [15], [17]-[18]. This value, called degree of membership, quantifies the grade of membership of the element in X to the fuzzy interval A.

Membership functions allow us to graphically represent a fuzzy set. The *y* axis represents the degrees of membership in the <0, 1> interval.

Defining fuzzy concepts, using more complex functions does not add more precision so simple functions are used to build membership functions. .

Another variant of the membership function which can be used is trapezoidal function. Given a general shape of a trapezoidal membership function as follows:

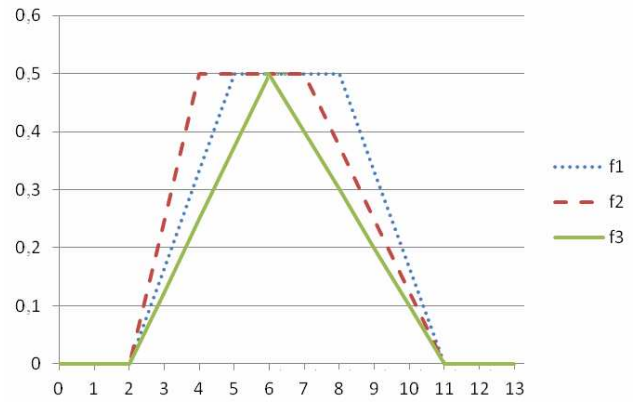

Figure 6. Examples of trapezoidal membership functions, where $a = 2$, $d = 11$, $y = 5$, $b = 4,5,6$ and $c = 7,8,9$; formula (7).

The parameters *a* and *d* can be specified by on the same formulas as it is for the triangular membership function, formula (6). According to *b* and *c* parameters, they based on the lower and upper quartil

$$f_A = \begin{cases} 0 & , x \le a \\ y(x-a)/(b-a), x \in (a,b] \\ y & , x \in (b,c] \\ y(d-x)/(d-c), x \in (c,d] \\ 0 & , \quad x > d \end{cases} \qquad (7)$$

The lower quartile b is the value of the middle of the first set of the response variable FATALR, where 25% of the values are smaller than b and 75% are larger. The upper quartile c is the value of the middle of the second set of the response variable FATALR, where 75% of the values are smaller than c and 25% are larger.

The difference between upper and lower quartiles (c-b, the interquartile range), indicates the dispersion of the response variable FATALR. The interquartile range spans 50% of the response variable FATALR, and eliminates the influence of outliers because, in effect, the highest and lowest quarters are removed.

3. Conclusion

Fitness of developed stochastic model of fatality rate FATALR turned out to be not satisfying especially at the beginning of modelled curve. Because of data big dispersion in the beginning extent of analyzed data set it was difficult to develop well fitted mathematical model. This is the reason of fuzzy regression usage. Such approach helped achieve better model adjustment to real data. Models of minimum and maximum values dependent on national product per capita NPPC were developed to determine fuzzy borders of FATALR variable. Such approach allowed to prepare fuzzy regression model of triangular membership function. Size of fuzzy sets is inversely proportional to NPPC variable.

References

[1] Buckley, J.J. (2006). *Fuzzy Statistics*. Taipei , Wunan.

[2] Cheng, C.H. (1998). A new approach for ranking by distance method. *Fuzzy Sets and Systems 95*, 307-317.

[3] Commandeur, J.J.F., Bijleveld, F.D., Bergel-Hayat, R., Antoniou, C., Yannis, G., & Papadimitriou, E. (2012). On statistical inference in time series analysis of the evolution of road safety. *Accident; Anal. Prev*. 1–11.

[4] Diamond, P. (1988). Fuzzy least squares. *Information Sciences 46 (3)*, 141-157.

[5] Dubois, D. & Prade, H. (1978). Operations on fuzzy numbers. *International Journal of Systems Science* 9, 613-626.

[6] Hinkley, D.V. & Reid, N. (1991). *Statistical Theory and Modelling: In Honour of Sir David Cox Frs*. Chapman & Hall.

[7] Holló, P., Eksler, V., & Zukowska, J. (2010). Road safety performance indicators and their explanatory value: A critical view based on the experience of Central European countries. *Safety Science* Vol. 48, No. 9, 1142–1150.

[8] Jamroz, K. (2011). *Metoda zarządzania ryzykiem w inżynierii drogowej*. Gdynia.

[9] Jamroz, K. (2012). The impact of road network structure and mobility on the national traffic fatality rate. *Energy Efficient Transportation Networks, International Scientific Conference, Paris 2012*.

[10] Van Beeck, A.E., Mackenbach, E., Looman, J.P. & Kunst, C.W. (1991). Determinants of Traffic Accident Mortality in the Netherlands: A Geographical Analysis. *International Journal of Epidemiology* Vol. 20, No. 3, 698–706.

[11] Wachnicka, J. (2012). Modelling selected road safety measures at the regional level in Europe. *Journal of Polish Safety and Reliability Association* Vol. 3, No. 2, ISNN:2084-5316.

[12] Wachnicka, J. & Smolarek, L. (2012). The Multivariate Multilevel Analysis Of Different Regional Factors Impact on Road Safety in European Country Regions. *Journal of KONBIN* No 4 (24), 141-148.

[13] Wachnicka, J. & Smolarek, L. (2013). Model of multilevel stochastic analysis of road safety on regional level. *Reliability: Theory & Applications* Vol.8 No. 9, 39-48.

[14] Tanaka, H., Uejima, S. & Asai, K. (1982). Linear regression analysis with fuzzy model. *Transactions on Systems Man and Cybernetics* Vol.12, 903-907.

[15] Zadeh, L.A. (1965). Fuzzy sets. *Inf. Control* 8, 338-353

[16] Zimmermann, H.J. (1987). *Fuzzy Set, Decision Making and Expert System*. Kluwer, Boston.

[17] Zimmermann, H.J. (1991). *Fuzzy Set Theory and its Application* (2nd ed.), Kluwer, Boston.

[18] http://www.dma.fi.upm.es/java/fuzzy/fuzzyinf/funpert_en.htm