


Article

A Fuzzy-Statistical Tolerance Interval from Residuals of Crisp Linear Regression Models

Maryam Al-Kandari ^{1,*}, Kingsley Adjenughwure ²  and Kyriakos Papadopoulos ¹

¹ Department of Mathematics, Kuwait University, P.O. Box 5969, Safat 13060, Khaldiya City, Kuwait; kyriakos@sci.kuniv.edu.kw

² Department of Civil Engineering, Democritus University of Thrace, 67100 Xanthi, Greece; kadjenug@civil.duth.gr

* Correspondence: maryam@sci.kuniv.edu.kw

Received: 10 August 2020; Accepted: 21 August 2020; Published: 25 August 2020



Abstract: Linear regression is a simple but powerful tool for prediction. However, it still suffers from some deficiencies, which are related to the assumptions made when using a model like normality of residuals, uncorrelated errors, where the mean of residuals should be zero. Sometimes these assumptions are violated or partially violated, thereby leading to uncertainties or unreliability in the predictions. This paper introduces a new method to account for uncertainty in the residuals of a linear regression model. First, the error in the estimation of the dependent variable is calculated and transformed to a fuzzy number, and this fuzzy error is then added to the original crisp prediction, thereby resulting in a fuzzy prediction. The results are compared to a fuzzy linear regression with crisp input and fuzzy output, in terms of their ability to represent uncertainty in prediction.

Keywords: tolerance interval; fuzzy linear regression; crisp linear regression; fuzzy-statistics

1. Introduction

In classical linear regression models, assumptions like linearity, fixed independent variables, normality of residuals, uncorrelated errors are made to simplify model estimation procedures. Despite these assumptions, the results are often taken at face value with very little effort, to adequately represent the uncertainty in predictions made by the model. Uncertainties in linear regression models are often represented via confidence and prediction intervals, which may not be adequate, since only one interval is calculated; e.g., 95% confidence or the prediction interval. Fuzzy linear regression with crisp input can be used to better represent uncertainty in prediction. This fuzzy model, first proposed in [1], has been widely used as alternative to classical crisp linear regression models. Since then, there have been various modifications to the model to overcome certain limitations of the original model [2–5]. Although such fuzzy linear regression models can represent uncertainty, there has always been doubt in terms of their suitability for prediction of future values (see a discussion on this in [6]). The problem is that minimization of fuzziness of the model is done, such that model fits the available sample with a certain h -value. There is no connection with prediction of future values like those of classic regression. Recently, there has been an attempt by [7] to create fuzzy numbers from predictions made by classic linear regression models, without the need of optimization or assuming any fuzzy coefficient, or fuzzy input or output. In their work, confidence and prediction intervals from crisp linear regression are converted to fuzzy numbers, by superimposing intervals and deriving the equivalent membership functions using fuzzy estimators. In this paper, we use the technique that was introduced in [7], and we propose a new approach to fuzzify the outputs of crisp linear regression models, for use in the case of tolerance intervals that are required, instead of prediction intervals. In our proposed approach, we assume that errors are normally distributed and, as such, we can construct a tolerance interval

of normal distribution for the errors. This tolerance interval will contain at least a proportion of the errors, both those in the sample and those outside of the sample (future predictions). Using the method proposed in [8], we construct a fuzzy number by superimposing the tolerance intervals up to the mean error. We then use this fuzzy tolerance interval as a fuzzy estimate of the error in our model. This is similar to the error proposed in [9], which uses crisp coefficients, and estimates a fuzzy error using optimization. This is to avoid the issue of the increasing magnitude of spread with an increasing independent variable. To complete the process, we add the fuzzy tolerance error to our crisp estimate, thereby resulting in a fuzzy estimate. The advantage of the proposed approach is that all possible statistical errors in the model are represented in one interval. Statistical tolerance intervals are, by definition, different from prediction intervals, and they serve a different purpose. Tolerance intervals give the percentage of population coverage interval with some confidence level, while a prediction interval will give the coverage interval for a single prediction. The interpretation and calculation of both intervals are also different. This paper thus extends the work in [7,8], to produce fuzzy statistical tolerance intervals. The method proposed here is important for applications where a tolerance interval is needed, instead of a prediction interval. The tolerance interval covers both the confidence interval and the prediction interval. Thus, both in-sample, out-of-sample, and future errors are covered. It also uses the information that model errors are normally distributed, so the better the original crisp model, the better the fuzzy model. Finally, the proposed method produces a fuzzy output with only crisp input, and crisp parameters without the need for optimization.

2. Crisp Linear Regression

A linear regression model with n independent variables and one dependent variable can be written as:

$$Y_k = \alpha_0 + \alpha_1 X_{k1} + \alpha_2 X_{k2} + \dots + \alpha_n X_{kn} + \varepsilon_k$$

where Y_k is the dependent variable, $X_{ki}, i = 1, 2, \dots, n$ are the independent variables, $\alpha_0, \alpha_1, \dots, \alpha_n$ are the coefficients which need to be estimated, and ε_k is the random error of the model. The linear regression model assumes that errors are normally distributed with zero mean and constant variance, i.e., $\varepsilon_k \sim N(0, \sigma)$.

The method of least squares is the most common way of estimating the model parameters. The parameters are calculated as:

$$A = (X^T X)^{-1} X^T Y$$

where A is a vector of the parameters, X is a matrix of explanatory variables, and Y is a vector of the response variable. A thorough examination of the distribution of errors is usually done after model estimation, to check the validity of the model.

To account for uncertainty in prediction due to random errors, prediction and confidence intervals are easily constructed both for the estimated parameters and for the predicted response. The $(1 - \alpha)\%$ prediction intervals are given as follows:

$$\bar{Y}_k \pm t_{1-\frac{\alpha}{2}, K-p} \sqrt{MSE(1 + h_f)}$$

where X is a matrix of explanatory variables, \bar{Y} is the estimated response variable, K is the sample size, MSE is the estimate of the mean-squared error of the model, $t_{1-\frac{\alpha}{2}, K-p}$ is a t-distribution with $K-p$ degrees of freedom, and $h_f = x_f (X^T X)^{-1} x_f$, where x_f is the row-vector of that observation.

3. Fuzzy Linear Regression

The classical fuzzy linear regression model proposed by [1] is similar to the crisp linear regression. The main difference is that the parameters are fuzzy numbers. This results in a fuzzy output for the response variable. The model is given below:

$$\widetilde{Y}_k = \widetilde{A}_0 + \widetilde{A}_1 X_{k1} + \widetilde{A}_2 X_{k2} + \dots + \widetilde{A}_n X_{kn}$$

where $\widetilde{A}_i (i = 1, 2, \dots, n)$ are symmetric triangular fuzzy numbers of the form (c_i, r_i) , c_i is the center of the triangular fuzzy number and r_i is the spread. The objective of the fuzzy linear regression model is to minimize the uncertainty by minimizing the spreads of the fuzzy numbers. This results in the linear optimization problem below [1]:

$$J = mc_0 + \sum_{j=1}^K \sum_{i=1}^n c_i |x_{ij}|$$

with the following constraints:

$$y_j \geq r_0 + \sum_{i=1}^n r_i x_{ij} - (1-h) \left(c_0 + \sum_{i=1}^n c_i |x_{ij}| \right)$$

$$y_j \leq r_0 + \sum_{i=1}^n r_i x_{ij} + (1-h) \left(c_0 + \sum_{i=1}^n c_i |x_{ij}| \right)$$

where $c_i \geq 0, i = 1, 2, \dots, n$. The value $0 \leq h \leq 1$ represents the confidence level of the model and the membership value of all all responses in the sample should be at least h i.e., $\mu(y_j) \geq h$ for $j = 1, 2, \dots, K$.

4. Proposed Method

Suppose that we have estimated a linear regression model and checked its validity with the necessary residual plots. That is, we assume that the model is well calibrated, and the distribution of errors are approximately normal. From every observation, our model produces an error:

$$\varepsilon_k = y_k - \hat{y}_k$$

where y_k is the real value of the response variable and \hat{y}_k is the estimate from the model. We assume that all errors in the model come from a normal distribution, with an unknown mean and unknown standard deviation. To accommodate the uncertainty in our errors, we can construct confidence intervals for the mean, or standard deviation of errors. However, this does not give us a bound on all errors that the model can produce, but only a bound for the mean. Additionally, a prediction interval can only hold for a particular prediction and, thus, it is not valid for other predictions. To accommodate both sample errors and future prediction errors, we opt to use a tolerance interval to bound the errors; an interval which contains $p\%$ of all errors with confidence $\gamma\%$. To simplify calculations, we do not focus on tolerance intervals for Y given a particular X (see for example [10]). Rather, we focus on calculating a general tolerance interval of a random sample originating from a normal distribution. We treat all errors in our model estimation as a random sample and try to find a tolerance interval for such errors. There are various ways to calculate the tolerance interval of a sample from a normal distribution [11–13]. For simplicity we choose, the approximation offered by [13]. The tolerance interval is defined below:

$$\bar{x} \pm K_2 s$$

$$K_2 = z_{\frac{1+p}{2}} \sqrt{\frac{\nu(1 + \frac{1}{N})}{\chi_{1-\gamma, \nu}^2}}$$

where $\bar{x} = \bar{\varepsilon}$ is the sample mean of the errors, s the sample standard deviation of the errors, K_2 is the tolerance factor, $\chi_{1-\gamma, \nu}^2$ is the critical value of the chi-square distribution with degrees of freedom ν that is exceeded with probability γ and $z_{\frac{(1+p)}{2}}$ is the critical value of the normal distribution associated with cumulative probability $\frac{1+p}{2}$.

To represent more proportions, we convert this interval to a fuzzy interval by superimposing all proportions up to $p = 0\%$, while the confidence level $\gamma\%$ is kept constant. Using the method proposed in [14,15], which was generalized in [8], we convert this interval to a fuzzy number with explicit membership function. In [8], it was shown that any interval of the form $[\bar{x} - mf(\alpha) \quad \bar{x} + mf(\alpha)]$ can be converted to a fuzzy number with explicit membership function if an appropriate function $f(\alpha)$ is chosen, where m is a constant and $f(\alpha)$ a function of α -cut of the corresponding fuzzy number. It has been shown that the inverse cumulative distribution functions are proper candidates for $f(\alpha)$. Since $z_{\frac{1+p}{2}}$ is an inverse cumulative distribution function, the superimposed intervals can be converted to a fuzzy number.

Now, the tolerance interval $[\bar{x} - K_2s \quad \bar{x} + K_2s]$ can be written in the form:

$$[\bar{x} - mf(\beta) \quad \bar{x} + mf(\beta)]$$

where $m = s \sqrt{\frac{v(1+\frac{1}{N})}{\chi^2_{1-\gamma, v}}}$, $f(\beta) = z_{\frac{\beta}{2}}$, $z_{\frac{\beta}{2}} = \Phi^{-1}(1 - \frac{\beta}{2})$.

Note that we substitute $z_{\frac{(1+p)}{2}} = z_{\frac{\beta}{2}}$ where $\beta = 1 - p$ for convenience; for example, $p = 0.90$ is equivalent to $\beta = 0.1$. The tolerance interval is then written as:

$$[\bar{x} - K_2s \quad \bar{x} + K_2s] = [\bar{x} - mz_{\frac{\beta}{2}} \quad \bar{x} + mz_{\frac{\beta}{2}}]$$

The interval above is exactly the $(1 - \beta)\%$ tolerance interval with confidence $\gamma\%$.

With the above substitutions, the following membership function can be derived for the interval:

$$A(x) = \begin{cases} \frac{2}{1-\beta} \Phi\left(\frac{x-\bar{x}}{m}\right) - \frac{\beta}{1-\beta}, & \bar{x} - m\Phi^{-1}\left(1 - \frac{\beta}{2}\right) \leq x \\ \frac{2}{1-\beta} \Phi\left(\frac{\bar{x}-x}{m}\right) - \frac{\beta}{1-\beta}, & x \leq \bar{x} + m\Phi^{-1}\left(1 - \frac{\beta}{2}\right) \\ 0, & \text{otherwise} \end{cases}$$

The α -cut of the fuzzy number above is [8]:

$${}^{\alpha}A = [\bar{x} - mz_{h(\alpha)} \quad \bar{x} + mz_{h(\alpha)}]$$

where $z_{h(\alpha)} = \Phi^{-1}(1 - h(\alpha))$ is the inverse cumulative distribution function (cdf) of a normal distribution and $h(\alpha)$ is any monotonic non-decreasing function $h(\alpha) : (0, 1] \rightarrow [\frac{\beta}{2}, 0.5]$, $\beta \in (0, 1)$, where $h(\alpha) = (\frac{1}{2} - \frac{\beta}{2})\alpha + \frac{\beta}{2}$.

Constructing a fuzzy output from the crisp linear regression, using the fuzzy number constructed from the tolerance interval of the errors, a crisp prediction from a linear regression model can be converted to a fuzzy output by adding the fuzzy error. The fuzzy output can be then written as:

$$\widetilde{Y}_k = \alpha_0 + \alpha_1 X_{k1} + \alpha_2 X_{k2} + \dots + \alpha_n X_{kn} + \widetilde{\varepsilon}$$

where $\widetilde{\varepsilon}$ is the fuzzy error constructed from the tolerance interval using the procedure described above. The reason for having only one error and not a prediction specific error is that it considers all errors in the model and, thus, it gives more conservative estimates compared to a prediction interval. Note that this approach is similar to the fuzzy linear regression model proposed in [9] but the error that they propose there is a fuzzy error estimated from an optimization process, and is not based on any statistical interval like the current one. Additionally, our method is similar to the one proposed in [7], but there the authors do not use the tolerance interval for errors like we do here. In their approach they convert well-known confidence and prediction intervals from linear regression to fuzzy numbers, using the same procedure used here. Our model can be viewed as a fuzzy linear regression model

with a fuzzy error constructed from a statistical tolerance interval. This can be used in the case where a tolerance interval is of interest, rather than a prediction or confidence interval.

5. Case Study

In order to test the applicability of our fuzzy linear regression model, two datasets from the fuzzy literature are used [16] and two real datasets from the classical linear regression literature (car data, UCI machine learning repository [17], Hald Cement data [18,19]). We compare the results to those of classical linear regression, to fuzzy linear regression of [1] and the fuzzy prediction intervals proposed in [7]. The datasets from the fuzzy literature are shown in the Tables 1–3 below.

Table 1. First dataset from Liu and Chen 2013 [16].

x_1	29.50	31.30	37.60	39.90	39.90	40.30	41.50	43.60	45.70	47.80	49.50
x_2	79.99	75.63	69.25	62.75	64.66	63.09	61.51	60.07	58.22	58.43	60.57
y	133.60	137.63	147.86	196.76	220.53	223.25	233.19	265.67	335.16	411.29	460.68

Table 2. First dataset from Liu and Chen 2013 [16].

x_1	50.10	50.20	49.90	50.00	50.00	50.00	50.90	53.10	55.20
x_2	58.23	58.03	57.53	55.68	55.24	54.51	50.08	50.05	49.72
y	477.96	474.02	466.80	466.16	469.80	468.95	476.24	499.39	521.20

Table 3. Second dataset from Liu and Chen 2013 [16].

x	2	4	6	8	10	12	16	18
y	14	16	14	18	18	22	18	22

Note that the comparison with a classical fuzzy model, and fuzzy prediction interval proposed in [7] is done for clarity, rather than to compare predictive performance. As it is well-known, tolerance interval, prediction interval, and the spreads from a fuzzy linear regression have different interpretations and are used for different purposes. Therefore, the results are not directly comparable. However, all three models share some similarities. For example, they can measure how close the true value is to our predicted value using the membership function of the predicted value. In addition, they can measure how much uncertainty is in the model, by using the spreads of the predicted values. In [16], the credibility of a predicted value is measured by how close it is to the original value and also how precise it is (i.e., how small the spreads of the value). The credibility of a fuzzy predicted value is defined as [16]:

$$Z_i = \frac{\mu_{\tilde{y}_i}(y_i)}{\Delta_{\tilde{y}_i}}$$

where $\mu_{\tilde{y}_i}(y_i)$ is the membership value of the true value, y_i in the fuzzy predicted value \tilde{y}_i and $\Delta_{\tilde{y}_i}$ is the fuzziness of the prediction which is equivalent to the area of the fuzzy number, which is a symmetric triangular fuzzy number with height of 1. The area is just the spread of fuzzy number, i.e., the difference between the central value and the right or left value.

For a model with sample size K , the total credibility of the model is the sum of the credibility of the individual sample predictions. The total credibility is given by:

$$TC = \sum_{i=1}^K Z_i = \sum_{i=1}^K \frac{\mu_{\tilde{y}_i}(y_i)}{\Delta_{\tilde{y}_i}}$$

Tables 4–15 below show the total credibility of all three models on all two datasets. The comparison with fuzzy linear regression is made for the h-value with the maximum credibility (indicated with * in the tables). However, since the h-value is not comparable to a statistical confidence level, we also

show the lowest membership value of the true value in the fuzzy number, that is produced from both a prediction and tolerance interval. A visual comparison is also shown in Figures 1 and 2. Since every data now belongs to the interval with some membership value, it is possible to calculate the minimum membership value in the dataset; this is what is defined as the lowest membership value, and it gives an indication of how good the coverage interval is. A value of zero indicates that there is at least one data that does not belong to the interval, while a value greater than zero implies that all data belong to the interval.

Table 4. Fuzzy model (data from Liu and Chen, 2013 [16]).

Hvalue	A	Total Credibility
0.3148	$y^* = (-610.8251, 0.0000) + (19.1484, 0.0000)x_1 + (1.4018, 1.2301)x_2$	0.1371
0	$y = (-610.8233, 0.0000) + (19.1484, 0.0000)x_1 + (1.4017, 0.8429)x_2$	0.1082
0.5	$y = (-610.8240, 0.0000) + (19.1484, 0.0000)x_1 + (1.4017, 1.6857)x_2$	0.1271

Table 5. Linear Model based on the fuzzy prediction intervals (data from Liu and Chen, 2013 [16]).

Confidence Level	Linear Model Based on Fuzzy Prediction Intervals (Data from Liu and Chen, 2013)	Total Credibility	Lowest Membership Value
90%	$y = (-1590.9) + (29.6726)x_1 + (9.9906)x_2$	0.2153	0
95%	$y = (-1590.9) + (29.6726)x_1 + (9.9906)x_2$	0.2056	0.0169
99%	$y = (-1590.9) + (29.6726)x_1 + (9.9906)x_2$	0.1949	0.0566

Table 6. Linear model based on the fuzzy tolerance intervals (data from Liu and Chen, 2013 [16]).

Proportion	Linear Model Based on Fuzzy Tolerance Interval (Data from Liu and Chen 2013)	Total Credibility	Lowest Membership Value
90%	$y = (-1590.9) + (29.6726)x_1 + (9.9906)x_2$	0.2029	0
95%	$y = (-1590.9) + (29.6726)x_1 + (9.9906)x_2$	0.1871	0.0793
99%	$y = (-1590.9) + (29.6726)x_1 + (9.9906)x_2$	0.1631	0.1794

Table 7. Fuzzy model (Data from Liu and Chen 2013 [16]).

Hvalue	Fuzzy Model (Data from Liu and Chen 2013)	Total Credibility
0.2666	$y^* = (12.0000, 1.3635) + (0.6250, 0.1704)x$	1.4960
0	$y = (12.0000, 1.000) + (0.6250, 0.1300)x$	1.2984
0.7	$y = (12.0000, 3.330) + (0.6250, 0.4200)x$	0.9736

Table 8. Linear Model based on the fuzzy prediction intervals (Data from Liu and Chen, 2013 [16]).

Confidence Level	Linear Model Based on Fuzzy Prediction Interval (Data from Liu and Chen 2013)	Total Credibility	Lowest Membership Value
90%	$y = (12.93) + (0.54)x$	1.5612	0.1371
95%	$y = (12.93) + (0.54)x$	1.4698	0.1825
99%	$y = (12.93) + (0.54)x$	1.3651	0.2155

Table 9. Linear model based on the fuzzy tolerance intervals (data from Liu and Chen, 2013 [16]).

Proportion	Linear Model Based on Fuzzy Tolerance Interval (Data from Liu and Chen 2013)	Total Credibility	Lowest Membership Value
90%	$y = (12.93) + (0.54)x$	1.3813	0.2921
95%	$y = (12.93) + (0.54)x$	1.1850	0.4034
99%	$y = (12.93) + (0.54)x$	0.8866	0.5627

Table 10. Fuzzy model (Car data, UCI repository).

Hvalue	Fuzzy Model (Car Data UCI Repository)	Total Credibility
0.3	$y = (45.4865, 13.8711) + (-0.0026, 0.0000)x_1 + (-1.091, 0.0000)x_2$	4.4629
0	$y^* = (45.4865, 9.7097) + (-0.0026, 0.0000)x_1 + (-1.091, 0.0000)x_2$	5.4725
0.5	$y = (45.4865, 19.4195) + (-0.0026, 0.0000)x_1 + (-1.091, 0.0000)x_2$	3.7626

Table 11. Linear model based on the fuzzy prediction intervals (Car data, UCI Repository).

Confidence Level	Linear Model Based on Fuzzy Prediction Intervals (Car data UCI Repository)	Total Credibility	Lowest Membership Value
90%	$y = (47.7694) + (-0.0066)x_1 + (-0.0420)x_2$	7.9395	0
95%	$y = (47.7694) + (-0.0066)x_1 + (-0.0420)x_2$	7.6792	0
99%	$y = (47.7694) + (-0.0066)x_1 + (-0.0420)x_2$	7.3592	0

Table 12. Linear Model based on the fuzzy tolerance intervals (Car data, UCI Repository).

Proportion	Linear Model Based on Fuzzy Tolerance Interval (Car data UCI Repository)	Total Credibility	Lowest Membership Value
90%	$y = (47.7694) + (-0.0066)x_1 + (-0.0420)x_2$	7.7065	0
95%	$y = (47.7694) + (-0.0066)x_1 + (-0.0420)x_2$	7.4172	0
99%	$y = (47.7694) + (-0.0066)x_1 + (-0.0420)x_2$	7.0796	0

Table 13. Fuzzy model (Hald Cement dataset [17]).

Hvalue	Fuzzy Model (Hald Cement Dataset [17])	Total Credibility
0.2487	$y^* = (79.3955, 0.0000) + (1.5212, 0.2474)x_1 + (0.3238, 0.0000)x_2 + (-0.0839, 0.1453)x_3 + (0.3187, 0.0000)x_4$	1.9208
0	$y = (79.3955, 0.0000) + (1.5212, 0.1859)x_1 + (0.3238, 0.0000)x_2 + (-0.0839, 0.1091)x_3 + (0.3187, 0.0000)x_4$	1.7102
0.5	$y^* = (79.3955, 0.0000) + (1.5212, 0.3718)x_1 + (0.3238, 0.0000)x_2 + (-0.0839, 0.2183)x_3 + (0.3187, 0.0000)x_4$	1.7059

Table 14. Linear model based on the fuzzy prediction intervals (Hald Cement dataset [17]).

Confidence Level	Linear Model Based on Fuzzy Prediction Intervals (Hald Cement Dataset [17])	Total Credibility	Lowest Membership Value
90%	$y = (62.4054) + (1.5511)x_1 + (0.5102)x_2 + (0.1091)x_3 + (-0.1441)x_4$	1.8326	0.0762
95%	$y = (62.4054) + (1.5511)x_1 + (0.5102)x_2 + (0.1091)x_3 + (-0.1441)x_4$	1.7302	0.1248
99%	$y = (62.4054) + (1.5511)x_1 + (0.5102)x_2 + (0.1091)x_3 + (-0.1441)x_4$	1.6160	0.1602

Table 15. Linear model based on the fuzzy tolerance intervals (Hald Cement dataset [17]).

Proportion	Linear Model Based on Fuzzy Tolerance Interval (Hald Cement Dataset [17])	Total Credibility	Lowest Membership Value
90%	$y = (62.4054) + (1.5511)x_1 + (0.5102)x_2 + (0.1091)x_3 + (-0.1441)x_4$	1.7655	0.1870
95%	$y = (62.4054) + (1.5511)x_1 + (0.5102)x_2 + (0.1091)x_3 + (-0.1441)x_4$	1.6683	0.2298
99%	$y = (62.4054) + (1.5511)x_1 + (0.5102)x_2 + (0.1091)x_3 + (-0.1441)x_4$	1.5724	0.2609

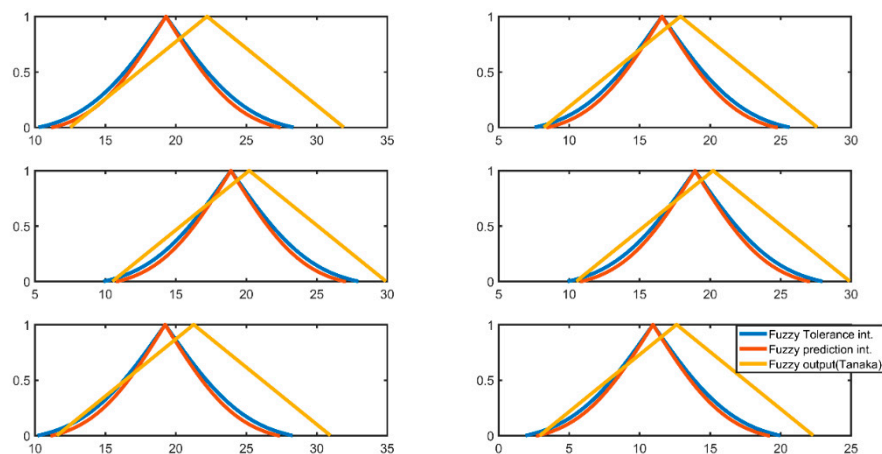


Figure 1. Some fuzzy predictions for fuel consumption in (miles per gallon) for the Car dataset, UCI Repository [17].

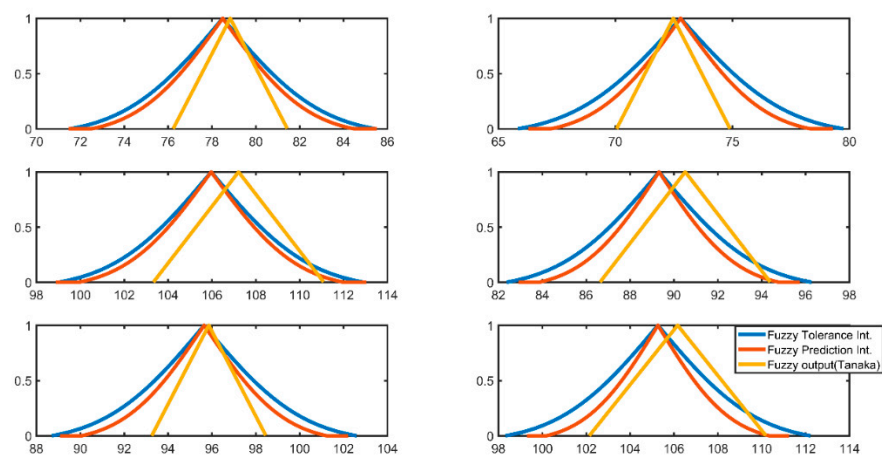


Figure 2. Some fuzzy predictions of heat evolved in (calories per gram) for the Hald Cement dataset [18].

The credibility of the proposed model for this dataset is slightly higher than that of the classical fuzzy model and lower than that of the fuzzy prediction intervals. As it can be seen, the credibility of all three models is within the same range of values, and this confirms the applicability of the proposed model. As expected, the credibility falls as the coverage probability increases. The 95% coverage seems to be the best coverage, since it contains all true values and a high credibility.

As with the previous dataset, the credibility of the proposed model is within the same range as the classical fuzzy model, and the fuzzy prediction interval further confirms its applicability. However, the values for both 90%, 95% and 99% coverage are slightly lower than the other two models. Again, the credibility falls as the coverage increases. All the three intervals contain the true values and with high membership values. The 90% coverage seems to give the best trade-off between total credibility and lowest membership value of true value.

6. Conclusions

Errors from a linear regression model are assumed to be normally distributed, with a mean of zero and constant standard deviation. In this study, we have shown how to construct a fuzzy error from the tolerance interval of the errors in a linear model. This approach leads to a fuzzy linear regression model with a constant fuzzy error. The fuzzy error is constructed entirely from the observed errors and it does not need to be optimized like the usual fuzzy regression models. The current model is useful for applications where a tolerance interval is needed to bound acceptable errors, in both sample and future predictions of a linear regression model.

For clarification, our method does not remove uncertainties in regression models, however, it gives a better method to use a fuzzy number to represent these uncertainties, instead of a single number or a single interval like it, as is normally done. In addition, choosing only one coverage interval, say 99%, it does not fully represent all the uncertainty in the model. In contrast, the fuzzy number produced contains all the uncertainty from, for example, 99% up to 0%.

In addition, by using a fuzzy number, another level of uncertainty is captured. Whether a number belongs to an interval or not is no longer crisp (0 or 1), but now one can say that a number belongs to the interval with some possibility [20], say 0.7. This captures the uncertainty in the assumptions that the interval will contain, say, 99% of the data. The fuzzy number shows that some data are more likely to be captured in the interval, compared to others.

Unlike crisp predictions, the true values can belong to the fuzzy output with a membership degree. As an added advantage, the total credibility of the model can be used to select among two linear regression models, and to choose the most credible model in terms of tolerance intervals for predictions. Additionally, similar to how h is used in a classical fuzzy model, a decision-maker can choose a linear model whose true value belongs to the fuzzy predicted values, with at least a specified membership degree.

The limitation of the model is that the fuzzy error is constant and is added to all crisp output. So, all outputs are assumed to be affected by the same error. In future research, non-constant tolerance interval could be used, which would depend on the input. The same procedure, for constructing a fuzzy number from a statistical interval, can be used to convert the tolerance intervals to fuzzy numbers. Last, but not least, different approximations exist for tolerance intervals, and they can be used apart from the one used in this paper. It would be worthwhile to extend this method to non-parametric tolerance intervals for solving the problem of assuming a known distribution.

Author Contributions: Investigation, K.A.; Methodology, K.A. and M.A.-K.; Resources, K.A. and K.P.; Supervision, K.A.; Validation, K.P.; Writing—original draft, M.A.-K., K.A. and K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tanaka, H.; Uegima, S.; Asai, K. Linear regression analysis with fuzzy model. *IEEE Trans. Syst. Man Cybernet.* **1982**, *12*, 903–907.
2. Tanaka, H.; Watada, J. Possibilistic linear systems and their application to the linear regression model. *Fuzzy Sets Syst.* **1988**, *27*, 275–289. [[CrossRef](#)]
3. Savić, D.; Pedrycz, W. Evaluation of fuzzy linear regression models. *Fuzzy Sets Syst.* **1991**, *39*, 51–63. [[CrossRef](#)]
4. Buckley, J.J. Fuzzy statistics: Hypothesis testing. *Soft Comput.* **2004**, *9*, 512–518. [[CrossRef](#)]
5. Chang, Y.-H.O.; Ayyub, B.M. Fuzzy regression methods—A comparative assessment. *Fuzzy Sets Syst.* **2001**, *119*, 187–203. [[CrossRef](#)]
6. Kim, K.J.; Moskowitz, H.; Köksalan, M. Fuzzy versus statistical linear regression. *Eur. J. Oper. Res.* **1996**, *92*, 417–434. [[CrossRef](#)]
7. Adjenughwure, K.; Papadopoulos, B. Constructing fuzzy-statistical prediction intervals from crisp linear regression models. In Proceedings of the 16th International Conference of Numerical Analysis and Applied Mathematics, Rhodes, Greece, 13–18 September 2018.
8. Adjenughwure, K.; Papadopoulos, B. Constructing fuzzy numbers from arbitrary statistical intervals. In Proceedings of the 2018 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Rhodes, Greece, 25–27 May 2018; pp. 1–6.
9. Kao, C.; Chyu, C.-L. A fuzzy linear regression model with better explanatory power. *Fuzzy Sets Syst.* **2002**, *126*, 401–409. [[CrossRef](#)]

10. Wallis, W.A. Tolerance intervals for linear regression. In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 31 July–12 August 1950; University of California Press: Berkeley, CA, USA, 1951.
11. Young, D.S. tolerance: An R Package for Estimating Tolerance Intervals. *J. Stat. Softw.* **2010**, *36*. [\[CrossRef\]](#)
12. Wald, A.; Wolfowitz, J. Tolerance Limits for a Normal Distribution. *Ann. Math. Stat.* **1946**, *17*, 208–215. [\[CrossRef\]](#)
13. Howe, W.G. Two-sided Tolerance Limits for Normal Populations—Some Improvements. *J. Am. Stat. Assoc.* **1969**, *64*, 610–620.
14. Sfiris, D.; Papadopoulos, B. Non-asymptotic fuzzy estimators based on confidence intervals. *Inf. Sci.* **2014**, *279*, 446–459. [\[CrossRef\]](#)
15. Dubois, D.; Foulloy, L.; Mauris, G.; Prade, H. Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities. *Reliab. Comput.* **2004**, *10*, 273–297. [\[CrossRef\]](#)
16. Liu, X.; Chen, Y. A systematic approach to optimizing h value for fuzzy linear regression with symmetric triangular fuzzy numbers. *Math. Probl. Eng.* **2013**. [\[CrossRef\]](#)
17. Lichman, M. *UCI Machine Learning Repository*; School of Information and Computer Science, University of California: Irvine, CA, USA, 2013. Available online: <http://archive.ics.uci.edu/ml> (accessed on 24 August 2020).
18. Woods, H.; Steinour, H.H.; Starke, H.R. Effect of composition of Portland cement on heat evolved during hardening. *Ind. Eng. Chem.* **1932**, *24*, 1207–1214. [\[CrossRef\]](#)
19. Montgomery, D.C.; Peck, E.A. *Introduction to Linear Regression Analysis*, 2nd ed.; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 1992.
20. Adjenughwure, K.; Papadopoulos, B. Fuzzy-statistical prediction intervals from crisp regression models. *Evol. Syst.* **2019**, *11*, 201–213. [\[CrossRef\]](#)



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).