Local linear regression analysis for interval-valued data

Jungteak Jang[a], Kee-Hoon Kang[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies, Korea

Abstract

Interval-valued data, a type of symbolic data, is given as an interval in which the observation object is not a single value. It can also occur frequently in the process of aggregating large databases into a form that is easy to manage. Various regression methods for interval-valued data have been proposed relatively recently. In this paper, we introduce a nonparametric regression model using the kernel function and a nonlinear regression model for the interval-valued data. We also propose applying the local linear regression model, one of the nonparametric methods, to the interval-valued data. Simulations based on several distributions of the center point and the range are conducted using each of the methods presented in this paper. Various conditions confirm that the performance of the proposed local linear estimator is better than the others.

Keywords: cross-validation, kernel function, nonlinear regression, nonparametric regression

1. Introduction

Regression analysis models how the average trend of response variable Y changes for a given explanatory variable X. Most data in this case correspond to single-valued data with only one fixed observation value. However, there are data that are difficult to represent by a single value because the observation objects have an internal structure and variation, and these are called symbolic data. Representative examples of symbolic data include multiple-valued data, histogram-valued data, and interval-valued data. Interval-valued data can be expressed as (1.1) in the form of intervals where the observed values of the variables have lower and upper limit values.

$$x_i = [x_{Li}, x_{Ui}]. \tag{1.1}$$

Here, x_{Li} and x_{Ui} denote the lower and upper limit value of the ith observation of the variable X, respectively.

Linear regression analysis of interval-valued data began to be studied in the early 2000s. References include Billard and Diday (2000, 2002), Lima Neto *et al.* (2004), Lima Neto and De Carvalho (2010), Xu (2010), Ahn *et al.* (2012), Im and Kang (2018), and so on. Many studies have been done on linear regression; however, there is limited research on nonlinear and nonparametric regression of interval-valued data.

Fagundes *et al.* (2014) proposed an interval kernel regression model that uses the Nadaraya-Watson type nonparametric regression estimator (IKR_NW) for the center point and range values of the intervals. An interval regression mixture model is also proposed therein, taking into account that center points and ranges can be represented as independent models. In this model, there is a model

---

which applies a kernel function to the center point and linear model to the range. However, there is another model that applies a linear model to the center point and a kernel function to the range. Recently, Lima Neto and De Carvalho (2017) proposed a nonlinear regression method (NLM) for interval-valued data.

In this paper, we review a nonparametric regression and a nonlinear regression model that focuses on interval-valued data. In addition, we propose a different version of nonparametric regression that uses local linear estimation. We compare the properties of the estimates addressed in this paper through various numerical simulations. The rest of the paper is organized as follows. In Section 2 we describe the interval-valued regression method using a Nadaraya-Watson type estimator. We also propose an approach that uses local linear regression to the center point and range value. The nonlinear regression model for interval-valued data is introduced in Section 3. In Section 4, we compare these methods through simulations. Finally, we draw a brief conclusion in Section 5.

2. Regression model using kernel function

Unlike linear regression analysis, nonparametric regression analysis is a technique to estimate the conditional expectation of response variable Y without assuming a distributional assumption when the explanatory variable X is given. A model like equation (2.1) can be considered by finding a functional relationship between response Y and explanatory variable X.

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.1}$$

Here, the function m is an unknown regression function, and ϵ_i's are identically and independently distributed with mean 0 and variance σ^2. A common method for estimating unknown function m is by using a kernel function. For given data (x_i, y_i), $i = 1, \ldots, n$, the kernel regression estimator proposed by Nadaraya (1964) and Watson (1964) is given by equation (2.2).

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}, \tag{2.2}$$

where K is a kernel function with $K_h(\cdot) = h^{-1} K(\cdot/h)$ and h is the smoothing parameter, which represents the smoothness of the estimated regression function.

2.1. Interval Kernel regression: Nadaraya-Watson estimator

It is possible to apply the kernel regression model to interval-valued data. Fagundes *et al.* (2014) proposed an interval kernel regression model that applies the Nadaraya-Watson regression estimator to the center point and range, respectively.

Assume that we have n observations $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$ from a response variable Y and p exploratory variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^t$. Here, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^t$, $x_{ij} = [x_{Lij}, x_{Uij}]$, $y_i = [y_{Li}, y_{Ui}]$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$. If we define a vector composed of center points of interval-valued data in an explanatory variable as $\boldsymbol{x}^c = (x_1^c, x_2^c, \ldots, x_p^c)^t$, the matrix consisting of the center points of each observation can be expressed as equation (2.3). The regression function of the center point can be estimated using the center point of the response variable and the matrices $\boldsymbol{X}^c$, where

$$\boldsymbol{X}^c = \left(X_1^c, X_2^c, \ldots, X_n^c \right), \quad X_i^c = \left(X_{i1}^c, \ldots, X_{ip}^c \right)^t, \quad i = 1, 2, \ldots, n. \tag{2.3}$$

Similarly, we can define a vector composed of range points as $\boldsymbol{x}^r = (x_1^r, x_2^r, \ldots, x_p^r)^t$ and others such as $\boldsymbol{X}^r = (X_1^r, X_2^r, \ldots, X_n^r)$, $X_i^r = (X_{i1}^r, \ldots, X_{ip}^r)^t$, $i = 1, 2, \ldots, n$.

Therefore, the relationship between the response variable and the explanatory variable is expressed as a combination of the center and the range, which can be in the form of equation (2.4).

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = [m^c(\boldsymbol{x}^c) - 0.5m^r(\boldsymbol{x}^r), m^c(\boldsymbol{x}^c) + 0.5m^r(\boldsymbol{x}^r)],\qquad (2.4)$$

where m^c is a function of the center point and m^r is a function of the range value. The estimated values of m^c and m^r are based on the Nadaraya-Watson estimator, as shown in (2.5).

$$\hat{m}^c(\boldsymbol{x}^c) = \sum_{i=1}^{n} w_i^c y_i^c, \quad \hat{m}^r(\boldsymbol{x}^r) = \sum_{i=1}^{n} w_i^r y_i^r,\qquad (2.5)$$

where

$$w_i^c = \frac{K_h\left(d\left(\boldsymbol{x}^c, X_i^c\right)\right)}{\sum_{i=1}^{n} K_h\left(d\left(\boldsymbol{x}^c, X_i^c\right)\right)}, \quad w_i^r = \frac{K_h\left(d\left(\boldsymbol{x}^r, X_i^r\right)\right)}{\sum_{i=1}^{n} K_h\left(d\left(\boldsymbol{x}^r, X_i^r\right)\right)}.$$

Here, we use the Gaussian kernel function K and bandwidth h given by

$$K_h\left(d\left(\boldsymbol{x}^c, X_i^c\right)\right) = \frac{1}{\left(\sqrt{2\pi}\right)^{\frac{1}{p}}} \frac{1}{h^p} e^{-\frac{d\left(\boldsymbol{x}^c, X_i^c\right)}{2h^2}},$$

where $d(\cdot)$ is the square of Euclidean distance between the two vectors.

2.2. Interval Kernel regression: local linear estimator

The Nadaraya-Watson estimator is known to have problems at the boundary of the data; however, local linear regression can overcome these drawbacks. With regard to nonparametric regression, it is natural to consider the local linear regression as well as the Nadaraya-Watson estimator. A multivariate local linear polynomial is of the form $\beta_0 + \boldsymbol{\beta}_1^t \boldsymbol{x}$, a hyperplane depending on $p + 1$ parameters β_0 and $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1p})^t$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^t$. Let K be a bounded, compactly supported p-variate kernel function satisfying

$$\int K(\boldsymbol{x})d\boldsymbol{x} = 1, \quad \int \boldsymbol{x}K(\boldsymbol{x})d\boldsymbol{x} = 0, \quad \text{and} \quad \int \boldsymbol{x}\boldsymbol{x}^t K(\boldsymbol{x})d\boldsymbol{x} = \mu_2(K)\boldsymbol{I},$$

where $\mu_2(K) = \int x_i^2 K(\boldsymbol{x})d\boldsymbol{x}$ is independent of i.

Let $\boldsymbol{H}$ be a bandwidth matrix which satisfies some regularity conditions. The multivariate version of local linear regression estimator is

$$\hat{m}(\boldsymbol{x}, 1, \boldsymbol{H}) = \boldsymbol{e}_1^t (\boldsymbol{X_x}^t \boldsymbol{W_x} \boldsymbol{X_x})^{-1} \boldsymbol{X_x}^t \boldsymbol{W_x} \boldsymbol{Y},$$

where

$$\boldsymbol{X_x} = \begin{bmatrix} 1 & (\boldsymbol{x} - \boldsymbol{X}_1)^t \\ \vdots & \vdots \\ 1 & (\boldsymbol{x} - \boldsymbol{X}_n)^t \end{bmatrix}$$

and

$$\boldsymbol{W_x} = \text{diag}\left\{K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_1), \ldots, K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_n)\right\}.$$

We can put the vector of center points or range points for each local linear regression estimator of center and range values, respectively.

In the case of a single explanatory variable, the local linear regression estimate is given in (2.6).

$$\hat{m}(x) = \hat{\beta}_0 = n^{-1} \sum_{i=1}^{n} \frac{\{\hat{s}_2(x;h) - \hat{s}_1(x;h)(x - x_i)\} K_h(x - x_i) y_i}{\hat{s}_2(x;h)\hat{s}_0(x;h) - \hat{s}_1(x;h)^2}, \tag{2.6}$$

where

$$\hat{s}_r(x;h) = n^{-1} \sum_{i=1}^{n} (x - x_i)^r K_h(x - x_i), \quad r = 0, 1, 2.$$

For each data center point and range, we can obtain the regression function estimates $\hat{m}^c$ and $\hat{m}^r$ by equation (2.6) and finally obtain the form of the interval by putting each estimate into equation (2.4).

3. Nonlinear regression model

Lima Neto and De Carvalho (2017) proposed the fit of a nonlinear regression model for interval-valued data. It is similar to the regression model for each of the previously proposed center points and ranges, but the difference is using the nonlinear function. Let y^c and y^r be the center point and range of the response variable, respectively, and let x_i^c and x_i^r be the center point and range of the p explanatory variables. When the unknown nonlinear regression functions for the center point and the range are f_c and f_r, the regression model is given by equation (3.1).

$$y_i^c = f_c\left(x_i^c, \theta^c\right) + \epsilon_i^c, \quad y_i^r = f_r\left(x_i^r, \theta^r\right) + \epsilon_i^r, \tag{3.1}$$

where ϵ_i^c, ϵ_i^r's are error terms with mean 0 and variance σ^2, θ^c, and θ^r are the parameter vectors for the model for center point and range.

Nonlinear regression functions f_c and f_r can be estimated by minimizing the residual sum of squares, as in (3.2).

$$\min_{\theta^c, \theta^r} \left\{ \sum_{i=1}^{n} \left(y_i^c - f_c\left(x_i^c, \theta^c\right)\right)^2 + \sum_{i=1}^{n} \left(y_i^r - f_r\left(x_i^c, \theta^r\right)\right)^2 \right\}. \tag{3.2}$$

To optimize this nonlinear objective function, we can use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, stochastic gradient, conjugate gradient, and simulated annealing method. In this paper, we employ the BFGS algorithm and conjugate gradient algorithm. See Edwin and Stanislaw (2013) for more detailed descriptions of these algorithms.

To estimate the interval values of the response variables, the predicted values $\hat{Y}^c$ and $\hat{Y}^r$ are obtained from the center-point model and the range model, respectively. And $\hat{Y} = [\hat{Y}_L, \hat{Y}_U]$ is obtained by using these estimates as in (3.3).

$$\hat{Y}_L = \min\left\{\hat{Y}^c - \frac{\hat{Y}^r}{2}, \hat{Y}^c + \frac{\hat{Y}^r}{2}\right\}, \quad \hat{Y}_U = \max\left\{\hat{Y}^c - \frac{\hat{Y}^r}{2}, \hat{Y}^c + \frac{\hat{Y}^r}{2}\right\}. \tag{3.3}$$

4. Simulation studies

Simulation was performed using open source statistical software R with version 3.5.1. We use **cvTools** package for cross-validation, **np** package for nonparametric regression and **optim** function to obtain
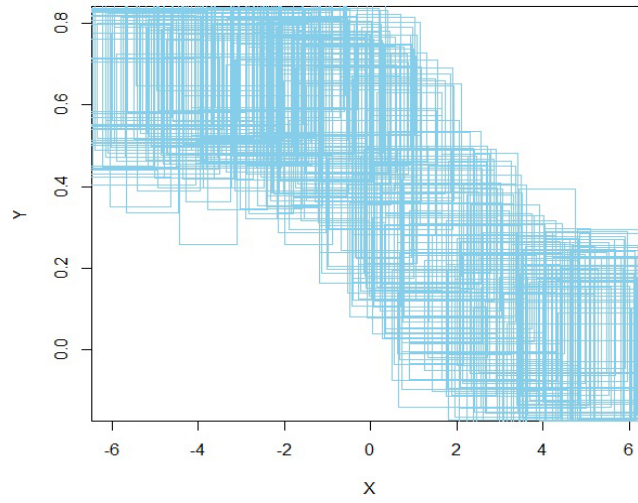
Figure 1: *A scatter plot of data set from the Experiment 1.*

parameter estimates in nonlinear regression models. Detailed usage and references can be found by using help in R.

In the experimental data, $n = 300$ samples were generated independently for the center point and the range. We use a Gaussian kernel function, and the smoothing parameter h is determined with the direct plug-in method proposed by Ruppert *et al.* (1995). For comparison, we used the lower boundary root mean squared error (RMSE_L) and the upper boundary root mean squared error (RMSE_U) in (4.1) proposed by Lima Neto and De Carvalho (2008). 10-fold cross-validation was applied separately to the modeling data and the validation data. The average values and their standard errors were calculated from experiments repeated 100 times.

$$\text{RMSE}_L = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{Li} - \hat{y}_{Li})^2} \quad \text{RMSE}_U = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{Ui} - \hat{y}_{Ui})^2}. \tag{4.1}$$

In this paper, the simulation is limited but considers types of data generation that can be inferred later. The simulation consists of two experiments and assumes a nonlinear form for both the center point and the range. Data sets for Experiment 1 in the simulation were generated as follows.

Experiment 1

Step 1. The center point and range of the explanatory variables x_i^c, x_i^r are generated from the following uniform distribution.

$$x_i^c \sim U(-6, 6), \quad x_i^r \sim U(1, 4).$$

Step 2. The errors for the center point and the range of the response variables are generated as follows.

$$\epsilon_i^c \sim N\left(0, 0.05^2\right), \quad \epsilon_i^r \sim N\left(0, 0.01^2\right).$$

Step 3. The parameters to be estimated are defined as follows.

$$\boldsymbol{\theta}^c = \left(\theta_1^c, \theta_2^c, \theta_3^c\right) = (2, 3, 1), \quad \boldsymbol{\theta}^r = (\theta_1^r, \theta_2^r) = (0.25, 1).$$

Table 1: Lower and upper boundary root mean squared errors in Experiment 1

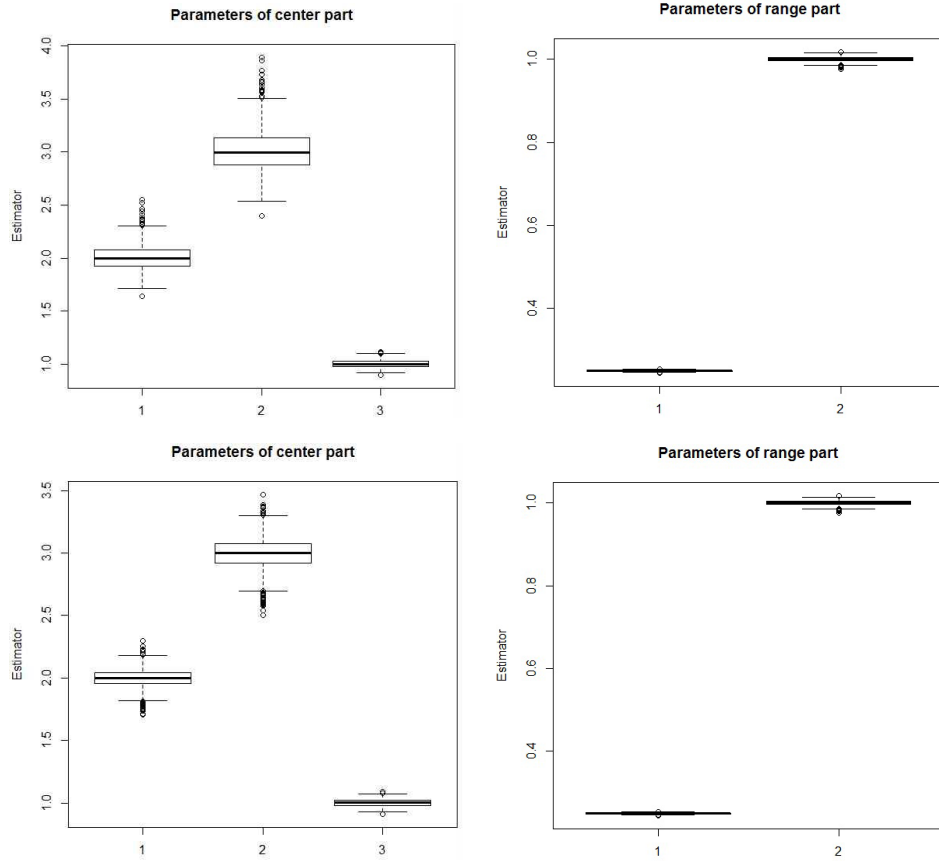| NLM_BFGS		NLM_CG		IKR_NW		IKR_LL	
$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$
0.0508	0.0507	0.0507	0.0506	0.0571	0.0570	0.0547	0.0544
(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)

Values in parentheses correspond to standard errors.

Figure 2: *Box plots of the parameter estimates from the nonlinear regression model in Experiment 1. The upper and lower panels correspond to the BFGS algorithm and CG algorithm, respectively. The left and right panels correspond to the center point and range, respectively.*

Step 4. The center point and range of the response variable are generated from the following model.

$$y_i^c = \frac{\theta_1^c}{\theta_2^c + \exp\left(\theta_3^c x_i^c\right)} + \epsilon_i^c, \quad y_i^r = \theta_1^r \exp\left(-\theta_2^r x_i^r\right) + \epsilon_i^r.$$

Figure 1 draws an example scatter plot of data set from Experiment 1. Table 1 and Figure 2 indicated the results of Experiment 1. NLM_BFGS and NLM_CG correspond to the nonlinear regression model estimated by using the BFGS algorithm and the conjugate gradient algorithm. IKR_NW
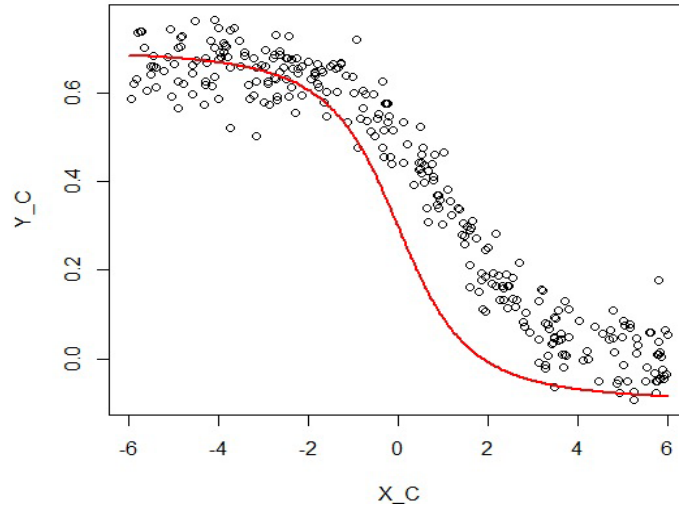
Figure 3: *A scatter plot of the center points for data from Experiment 1 and graph of the model (4.2) with* $(\theta_1^c, \theta_2^c, \theta_3^c, \theta_4^c) = (-0.8, 11, 4, 0.3)$.

Table 2: Results of applying incorrect form of nonlinear regression function in Example 1

NLM_BFGS		NLM_CG		IKR_NW		IKR_LL	
$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$
0.5069	0.5070	0.5164	0.5166	0.0571	0.0570	0.0547	0.0544
(0.0012)	(0.0011)	(0.0011)	(0.0011)	(0.0002)	(0.0002)	(0.0002)	(0.0002)

Values in parentheses correspond to standard errors.

and IKR_LL correspond to the kernel estimates by the Nadaraya-Watson method and the local linear method.

The root mean squared errors in Table 1 show that the errors of the nonlinear regression models are slightly lower than the ones of kernel methods. Among kernel methods the proposed local linear estimates show better performance than Nadaraya-Watson estimates. In addition, Figure 2 shows box plots of the parameter estimates for the center point and the range model. It can be seen that the parameters in the range are estimated more accurately than the ones in the center point. In Table 1 and Figure 2, the two nonlinear methods show similar results.

However, this result comes from assuming the shape of a true model for parameter estimation. In practice, it is not easy to consider the actual function by looking at a scatter plot of the interval-valued data. To investigate this situation, Figure 3 shows a scatter plot of the center point data and the nonlinear regression function that can be deduced from it in Experiment 1. The nonlinear regression function, which is a candidate for the center point, is drawn from the form of (4.2) with $\boldsymbol{\theta}^c = (\theta_1^c, \theta_2^c, \theta_3^c, \theta_4^c) = (-0.8, 11, 4, 0.3)$.

$$f(x_i^c, \boldsymbol{\theta^c}) = \frac{\theta_1^c x_i^c}{\sqrt{\theta_2^c + \theta_3^c (x_i^c)^2}} + \theta_4^c. \tag{4.2}$$

Under the center point model in (4.2), the results in Table 2 reveal that the nonlinear models show poor performance; consequently, the results of the kernel method remain unchanged. The result can
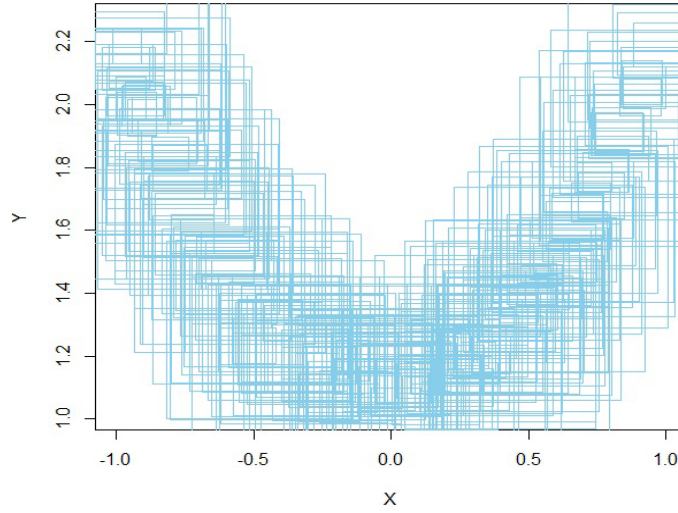
Figure 4: *A scatter plot of data set from the Experiment 2.*

therefore vary significantly depending on the correctness of the considered nonlinear function in the case of the nonlinear regression model.

We now introduce the next experiment. The process of generating the data sets of Experiment 2 is as follows; in addition, we include an example scatter plot drawn in Figure 4.

Experiment 2

Step 1. The center point and range of the explanatory variables x_i^c, x_i^r are generated from the following uniform distribution.

$$x_i^c \sim U(-1, 1), \quad x_i^r \sim U(0.1, 0.7).$$

Step 2. The errors for the center point and range of the response variables are generated as follows.

$$\epsilon_i^c \sim N\left(0, 0.05^2\right), \quad \epsilon_i^r \sim N\left(0, 0.01^2\right).$$

Step 3. The parameters to be estimated are defined as follows.

$$\boldsymbol{\theta}^c = \left(\theta_1^c, \theta_2^c\right) = (1, 3), \quad \boldsymbol{\theta}^r = (\theta_1^r, \theta_2^r) = (0.1, 1).$$

Step 4. The center point and range of the response variable are generated from the following model.

$$y_i^c = \left(\theta_1^c + \left(\frac{\theta_2^c}{\pi} x_i^c\right)^2\right)^{\frac{\pi}{\exp(1)}} + \epsilon_i^c, \quad y_i^r = \theta_1^r + \theta_2^r \left(x_i^r\right)^2 + \epsilon_i^r.$$

The results of Experiment 2 are presented in Figure 5 and Table 3. Figure 5 shows that the parameters for the center point in the nonlinear regression model using the BFGS algorithm are not well estimated, but are correctly estimated using the CG algorithm. One of the referees requested us to
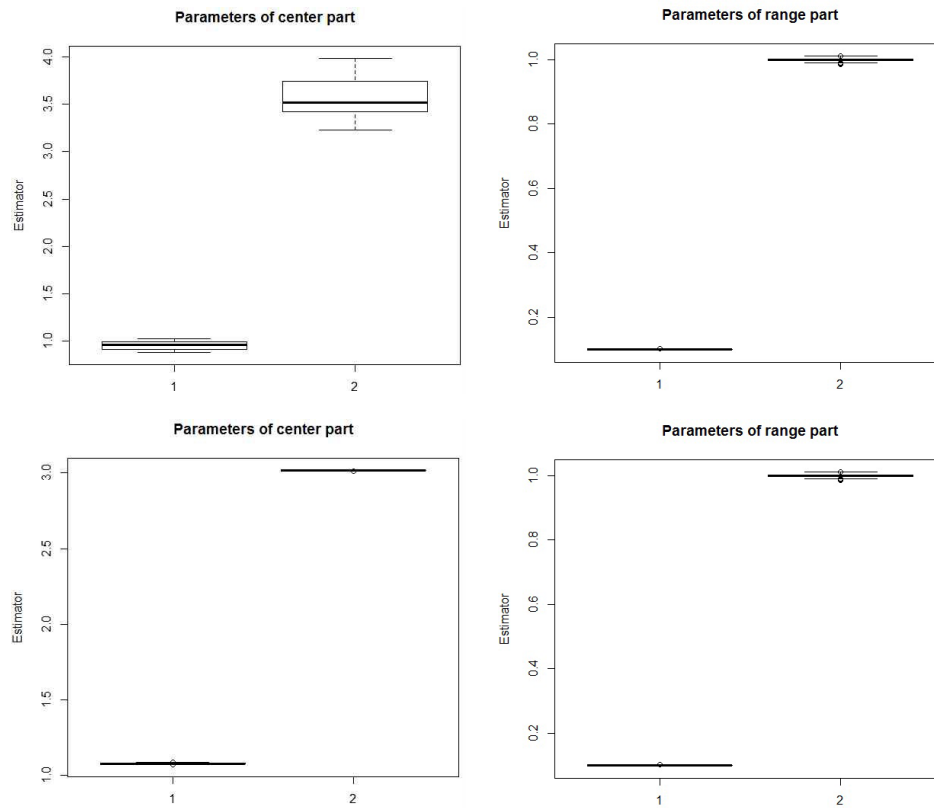
Figure 5: *Box plots of the parameter estimates from the nonlinear regression model in Experiment 2. The upper and lower panels correspond to the BFGS algorithm CG algorithm. The left and right panels correspond to the center point and range.*

Table 3: Lower and upper boundary root mean squared errors in Experiment 2

NLM_BFGS		NLM_CG		IKR_NW		IKR_LL	
$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$
0.1468	0.1468	0.0511	0.0508	0.1057	0.1057	0.0614	0.0637
(0.0034)	(0.0039)	(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0007)	(0.0007)

Values in parentheses correspond to standard errors.

explain why this happened, we are not able to give an accurate explanation. It should be noted, however, that Qin (2007) mentioned that CG is generally the best method for well-conditioned problems. Table 3 shows that the results might be bad if the parameters are not properly estimated in the non-linear regression model. The nonlinear method using the CG algorithm shows lower errors than the kernel regression cases. The nonlinear regression model shows lower errors only when the parameter is correctly estimated. Even in this case, it should be noted that the nonparametric method using local linear regression shows a relatively good performance.

However, it is difficult to guess the true form of the nonlinear regression function by only seeing a scatter plot of the interval-valued data. To see this situation, a scatter plot of the center point data and the nonlinear regression function that can be deduced from this plot in Experiment 2 are shown
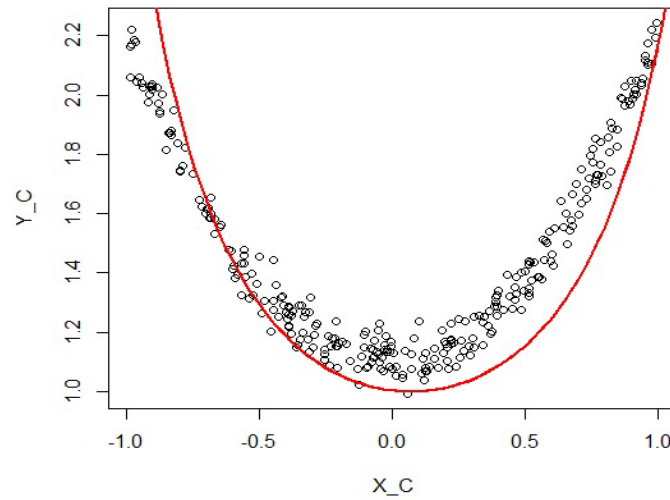
Figure 6: *A scatter plot of the center points for data from Experiment 2 and graph of the model (4.4) with* $\boldsymbol{\theta}^c = (\theta_1^c, \theta_2^c) = (1, 1.5)$.

Table 4: Results of applying incorrect form of nonlinear regression function in Example 2

NLM_BFGS		NLM_CG		IKR_NW		IKR_LL	
$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$	$RMSE_L$	$RMSE_U$
0.3571	0.3570	0.2986	0.2986	0.1057	0.1057	0.0614	0.0637
(0.0045)	(0.0046)	(0.0016)	(0.0016)	(0.0005)	(0.0005)	(0.0007)	(0.0007)

in Figure 6. Candidates for the nonlinear function of this center point regression can take the form of equation (4.3) with the parameter $\boldsymbol{\theta}^c = (\theta_1^c, \theta_2^c) = (1, 1.5)$.

$$f\left(x_i^c, \boldsymbol{\theta}^c\right) = \left(\sin\left(\theta_1^c x_i^c + \theta_2^c\right)\right)^{-1.5}. \tag{4.3}$$

In this case, the results in Table 4 show that nonparametric estimation using the kernel function is very good, and the IKR_LL model, which is a local linear regression method, is more excellent. The performance of the two nonlinear regression methods becomes drastically worse. Again, we see that the performance of the nonlinear regression model depends on the correctness of the proposed function.

According to comments by the referees, we have increased the variances of errors in both of Experiments 1 and 2. We do not include all results in this paper; however, we draw exactly the same conclusions as the previous ones. Detailed numerical results can be requested from the authors.

5. Conclusion

In this paper, we briefly reviewed the Nadaraya-Watson estimator and nonlinear regression model for analyzing interval-valued data. Additionally, we proposed a local linear regression estimator as one of the nonparametric approaches. This paper uses simulation without comparing the theoretical characteristics of the various methods. However, we admit that the general conclusion is limited because not all situations can be considered in the simulation.

The simulation results in this paper suggest that the local linear regression model shows better performance than the Nadaraya-Watson estimator. This is true for near linear and nonlinear assumptions about the distribution of each center and range model. We can use the BFGS algorithm and the conjugate gradient algorithm to estimate the nonlinear parameters in the model when nonlinear regression models are used to fit the regression function.

Based on this, the nonlinear regression model and nonparametric models were compared with each other under the assumption of a true nonlinear regression function. The NLM model performed better and calculation speed was not lower. However, the IKR_LL model showed relatively good performance even under the assumption of a true nonlinear function, and was superior when it was difficult to consider the true nonlinear function or if it was estimated incorrectly. Among the two NLM methods, the conjugate gradient algorithm gave better results than the BFGS algorithm.

Therefore, the NLM model can have better prediction power in the case of a nonlinear structure when the real function can be grasped easily by looking at a scatter plot of the interval data. However, the IKR_LL model has been found to be more predictable when it is difficult to estimate the actual nonlinear function only by the data structure.

Acknowledgements

References

Ahn J, Peng M, Park C, and Jeon Y (2012). A resampling approach for interval-valued data regression, *Statistical Analysis and Data Mining*, **5**, 336–348.

Billard L and Diday E (2000). Regression analysis for interval-valued data. In *Data Analysis, Classification, and Related Methods*, (H. A. L. Kiers, J.-P Rassoon, P. J. F. Groenen, and M. Schader (eds), pp. 369–374), Springer-Verlag, Berlin.

Billard L and Diday E (2002). Symbolic regression analysis. In *Classification, Clustering and Data Analysis*, (K. Jajuga, A. Sokolowski, and H.-H Bock (eds), pp. 281–288), Springer-Verlag, Berlin.

Edwin KPC and Stanislaw HZ (2013). *An Introduction to Optimization* (4th ed), Wiley, New Jersey.

Fagundes RAA, De Souza RMCR, and Cysneiros FJA (2014). Interval kernel regression, *Neurocomputing*, **128**, 371–388.

Im S and Kang K (2018). On regression analysis of interval-valued data, *Journal of the Korean Data & Information Science Society*, **29**, 351–365.

Lima Neto EA and De Carvalho FAT (2008). Center and range method for fitting and linear regression model to symbolic interval data, *Computational Statistics and Data Analysis*, **52**, 1500–1515.

Lima Neto EA and De Carvalho FAT (2010). Constrained linear regression models for symbolic interval-valued variables, *Computational Statistics and Data Analysis*, **54**, 333–347.

Lima Neto EA and De Carvalho FAT (2017). Nonlinear regression applied to interval-valued data, *Pattern Analysis and Applications*, **20**, 809–824.

Lima Neto EA, De Carvalho FAT, and Tenorio C (2004). Univariate and multivariate linear regression methods to predict interval-valued features. In *Advances in Artificial Intelligence 2004* (G.I. Webb and X. Yu (eds), Vol 3339, pp. 526–537), Springer, Berlin, Heidelberg.

Nadaraya EA (1964). On estimating regression, *Theory of Probability and Its Application*, **10**, 186–

190.

Qin Z (2007). The relationships between CG, BFGS, and two limited-memory algorithms, *Electronic Journal of Undergraduate Mathematics*, **12**, 5–20.

Ruppert D, Sheather SJ, and Wand MP (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, **90**, 1257–1270.

Watson GS (1964). Smooth regression analysis, *Sankhyā: The Indian Journal of Statistics, Series A*, **26**, 359–372.

Xu W (2010). *Symbolic data analysis: interval-valued data regression* (Ph.D, dissertation), University of Georgia, Georgia.