An Implementation of Fuzzy Linear Regression Module and its Application of Forecasting Electricity Demand

黃提道

Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan

Abstract

We implement a fuzzy linear regression module that can deal with quantitative variables and qualitative variables as well. Moreover, we modify the LP formulation behind the fuzzy regression and increase its fitting performance. To evaluate the performance of the fuzzy linear regression module, we conduct an empirical analysis of forecasting the electricity demand of a manufacturing factory in Shanghai. The predictors include production quantity, quarter (Q1, Q2, Q3, and Q4), temperature, precipitation, and wind speed. Also, electricity usage is taken as response. The results based on training set and testing set show that mean absolute percentage error (MAPE) of fuzzy linear regression almost equals to that of OLS, which implies that fuzzy linear regression is as good as conventional regression methodology. However, without additional statistical assumption under fuzzy linear regression, fuzzy linear regression is easy to perform. If the assumption of conventional linear regression cannot be satisfied, then fuzzy linear regression may be a good alternative for researchers to take advantage of.

Keywords: Fuzzy linear regression, electricity demand forecasting

I. Introduction

Fuzzy linear regression was proposed by Tanaka et al. [1], and its equation was:

$$Y = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \qquad \textbf{(1)}$$

where $\beta_j, j = 0,1,\ldots,p$, was the regression coefficient, assumed to be a symmetric triangular fuzzy number with center b_i and half-width w_i, $w_i \geq 0$ (see Fig. 1). The response y and predictorx_i, $i = 0,1,\ldots,p$, could be numbers (that is crisp values) or fuzzy numbers. In this paper, we only consider the case of fuzzy number of responses and crisp values of predictors. Moreover, the fuzzy number of response y is supposed to be a symmetric triangular number. In [2], Hojati et al. have introduced h-certain observed interval and h-certain predicted interval in case we sometimes are only interested in that part of y_i, $i = 1,\ldots,n$, where n represents the number of input data. The h-certain observed interval is written as $[y_i - (1-h)e_i, y_i + (1-h)e_i]$, where e_i represents the half-width of fuzzy number y_i, and h represents a minimal membership value allowed in the model, $0 \leq h \leq 1$ (see Fig.

2). Similarly, the h-certain predicted interval is expressed as $[\sum_{j=0}^{p}\left(\beta_j-(1-h)w_j\right)x_{ij}\,,\sum_{j=0}^{p}\left(\beta_j+(1-h)w_j\right)x_{ij}]$.

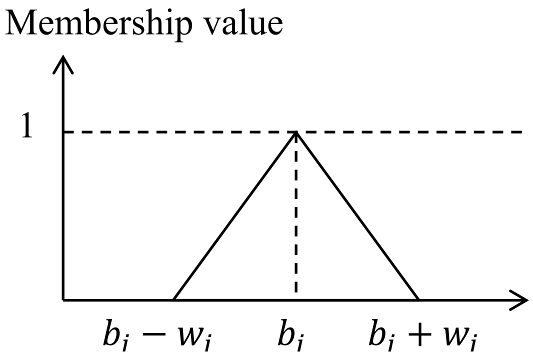


Fig. 1. A triangular fuzzy number

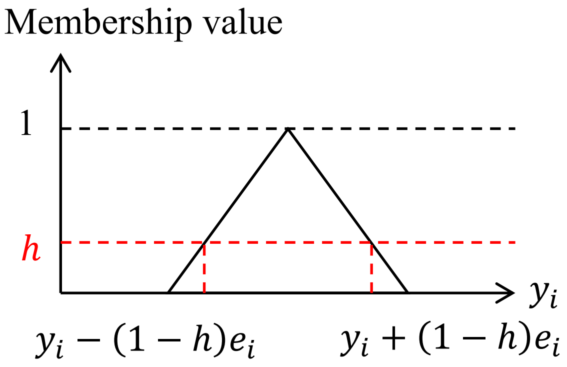


Fig. 2. A h-certain observed interval

To obtain the symmetric triangular coefficients, the most common methodology is to solve linear programming (LP) problem with the objective function of minimizing the sum of deviations of predicted response. The early LP model introduced by Tanaka et al. [1] to estimate $\beta_j, j=0,1,\dots,p$ is shown as follows:

$$\min \sum_{j=0}^{p} w_j \tag{2}$$

subject to

$$\sum_{j=0}^{p}\left(\beta_j+(1-h)w_j\right)x_{ij} \geq y_{ij}+(1-h)e_i, i=1,\dots,n, \tag{3}$$

$$\sum_{j=0}^{p}\left(\beta_j-(1-h)w_j\right)x_{ij} \leq y_{ij}-(1-h)e_i, i=1,\dots,n, \tag{4}$$

$$\beta_j=\text{free}, w_j\geq 0, j=0,\dots,p. \tag{5}$$

Tanaka et al's model is sensitive to the potential outlier of the response because the concept of this model is to force the h-certain predicted interval to cover all the h-certain observed interval. To deal with this setback, Hojati et al [2] have proposed a new LP formulation:

$$\min \sum_{i=1}^{n}(d_{iU}^{+}+d_{iU}^{-}+d_{iL}^{+}+d_{iL}^{-}) \tag{6}$$

subject to

$$\sum_{j=0}^{p}\left(\beta_j+(1-h)w_j\right)x_{ij}+d_{iU}^{+}-d_{iU}^{-}=y_{ij}+(1-h)e_i, i=1,\dots,n, \tag{7}$$

$$\sum_{j=0}^{p}\left(\beta_j-(1-h)w_j\right)x_{ij}+d_{iL}^{+}-d_{iL}^{-}=y_{ij}-(1-h)e_i, i=1,\dots,n, \tag{8}$$

$$\beta_j=\text{free}, w_j\geq 0, j=0,\dots,p, \tag{9}$$

where $|d_{iU}^{+}-d_{iU}^{-}|$ indicates the distance between upper bound of h-certain predicted interval and upper bound of h-certain observed interval, and $|d_{iL}^{+}-d_{iL}^{-}|$ indicates the distance between lower bound of h-certain predicted interval and lower bound of h-certain observed interval. The equation (7), coming from equation (3), adds 2 new variables d_{iU}^{+} and d_{iU}^{-} (surplus variable and artificial variable), which implies that only one of d_{iU}^{+} and d_{iU}^{-}would be positive. Furthermore, the equation (8), originated from equation (4), introduces 2 new variables d_{iL}^{+} and d_{iL}^{-}, which are both slack variables. Because of the duplicated slack variables, we know that only one of d_{iL}^{+} and d_{iL}^{-} will be positive, and another will be zero. Consequently, the objective function is equivalent to minimize the sum of $|d_{iU}^{+} - d_{iU}^{-}|$ and $|d_{iL}^{+} - d_{iL}^{-}|$, and this formulation has properly solved the setback of Tanaka et al's model.

The rest of the present work is organized as follows. In section II, we implement a fuzzy linear regression module with both Tanaka et al's and Hojati et al's formulations. In the fact that most of cases or publications related to fuzzy linear regression only consider the quantitative predictors [2-5], we introduce sum coding to deal with qualitative predictors; therefore, the proposed fuzzy linear regression can estimate the coefficients of quantitative predictors and qualitative predictors as well. In order to judge the functionality and performance of the fuzzy linear regression module, we perform a case study of electricity demand forecasting of the factory of a manufacturing company in Shanghai in section III, IV, and V. The conclusion is given in section VI.

II. Implementation

In this project, we use python to realize the whole process of fuzzy linear regression, and the detail of implementation is shown in Fig. 3.

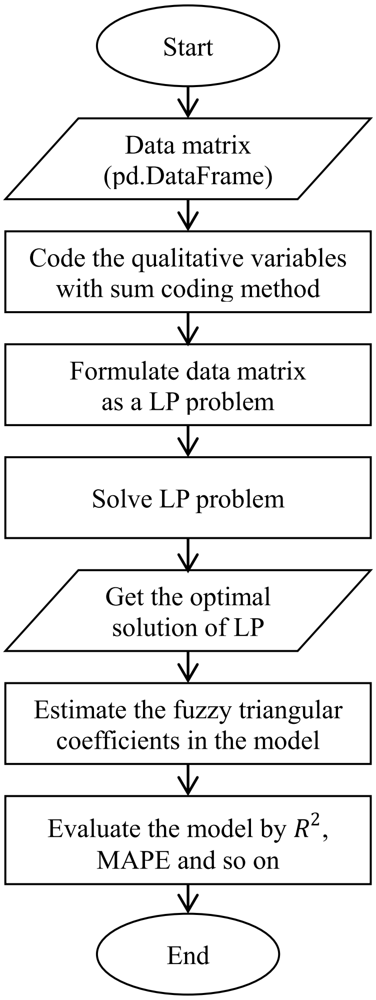
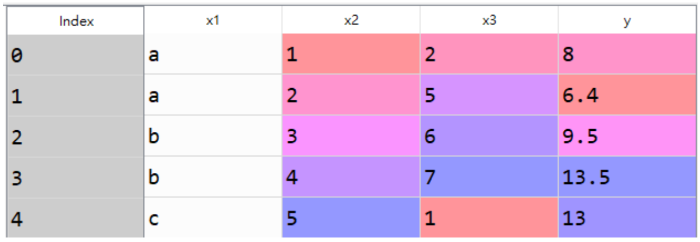

Start
Data matrix (pd.DataFrame)
Code the qualitative variables with sum coding method
Formulate data matrix as a LP problem
Solve LP problem
Get the optimal solution of LP
Estimate the fuzzy triangular coefficients in the model
Evaluate the model by R^2, MAPE and so on
End

Fig. 3. The procedure of implanting fuzzy linear regression model

Fig. 3 shows that the data matrix is in the format of DataFrame of pandas package, demonstrated in Fig. 4.



Index	x1	x2	x3	y
0	a	1	2	8
1	a	2	5	6.4
2	b	3	6	9.5
3	b	4	7	13.5
4	c	5	1	13

Fig. 4. An example of data matrix of fuzzy linear regression module

In Fig. 4, we may notice predictor x_1 is a qualitative variable; therefore, the module will automatically apply sum coding to variable x_1 so that the fuzzy linear regression model still works. In the process of formulation, we introduce LP formulation from Tanaka et al. [1] and Hojati et al [2], and name these two formulations 'Tan' and 'HBS' respectively. Recall the property of the conventional regression model:

$$Y = \alpha_0 x_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2), \qquad \textbf{(10)}$$

where $E(\varepsilon) = 0$ implies the property that the fitted regression line always goes through the point $(\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_p, \bar{y})$, which may enhance the fitting performance. However, Tan and HBS model do not satisfy the property. Therefore, the proposed fuzzy linear regression module provides an option for user to add a new linear constraint:

$$\bar{y} = \alpha_0\bar{x}_0 + \alpha_1\bar{x}_1 + \cdots + \alpha_p\bar{x}_p, \tag{11}$$

to be sure that the fuzzy linear regression model may possess the same property of the conventional regression model. In the procedure of solving LP we introduce the package of CPLEX Python API, created by IBM ILOG CPLEX Optimization Studio, to obtain the optimal solution of LP.

The last step, we use different criterions to evaluate the fuzzy regression model. The first criterion is R^2, defined as:

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{12}$$

The equation (12) is as same as R^2 of the conventional linear regression. However, the R^2 of the fuzzy regression does not hold the theorem of Sum-of-Square Identity (see the proof in Appendix A); therefore, this R^2 may not fall between 0 and 1. The second criterion is mean absolute percentage error (MAPE), which follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{y_i - \bar{y}}{y_i} \right|. \tag{13}$$

The last criterion is jaccard similarity measurement, which is calculated by:

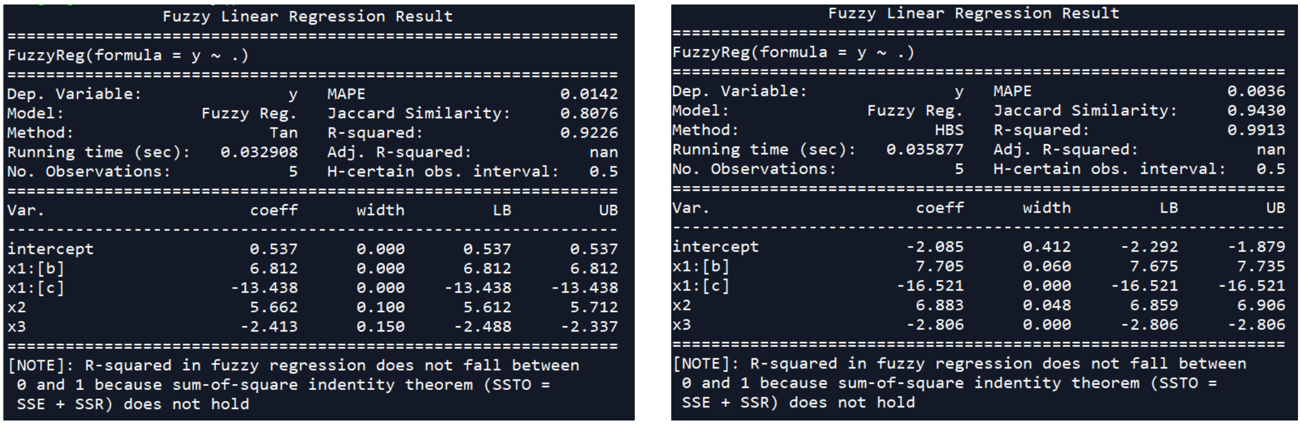
$$jaccard\ similarity = \frac{1}{n}\sum_{i=1}^{n}\frac{intersection_i}{predicted\ range_i}, \tag{14}$$

where $predicted\ range_i$ and $intersection_i$, $i = 1, \dots, n$, represent the range of h-certain predicted interval and the intersection between h-certain observed interval and h-certain predicted interval respectively.

In this section, we use the dataset shown in Fig, 4 to perform a fuzzy linear regression analysis. The conceptual model will be like:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i}, i = 1, \dots, n, \tag{15}$$

and we assume that response y_i is a fuzzy triangular number with the width of $\frac{y_i}{20}$, $i = 1, \dots, n$, and minimal membership value h is 0.5.



```
                 Fuzzy Linear Regression Result
==============================================================
FuzzyReg(formula = y ~ .)
==============================================================
Dep. Variable:                y   MAPE                  0.0142
Model:               Fuzzy Reg.   Jaccard Similarity:   0.8076
Method:                     Tan   R-squared:            0.9226
Running time (sec):    0.032908   Adj. R-squared:          nan
No. Observations:             5   H-certain obs. interval:   0.5
==============================================================
Var.                     coeff      width         LB        UB
--------------------------------------------------------------
intercept                0.537      0.000      0.537     0.537
x1:[b]                   6.812      0.000      6.812     6.812
x1:[c]                 -13.438      0.000    -13.438   -13.438
x2                       5.662      0.100      5.612     5.712
x3                      -2.413      0.150     -2.488    -2.337
==============================================================
[NOTE]: R-squared in fuzzy regression does not fall between
 0 and 1 because sum-of-square indentity theorem (SSTO =
 SSE + SSR) does not hold
```

(a) Tan

```
                 Fuzzy Linear Regression Result
==============================================================
FuzzyReg(formula = y ~ .)
==============================================================
Dep. Variable:                y   MAPE                  0.0036
Model:               Fuzzy Reg.   Jaccard Similarity:   0.9430
Method:                     HBS   R-squared:            0.9913
Running time (sec):    0.035877   Adj. R-squared:          nan
No. Observations:             5   H-certain obs. interval:   0.5
==============================================================
Var.                     coeff      width         LB        UB
--------------------------------------------------------------
intercept               -2.085      0.412     -2.292    -1.879
x1:[b]                   7.705      0.060      7.675     7.735
x1:[c]                 -16.521      0.000    -16.521   -16.521
x2                       6.883      0.048      6.859     6.906
x3                      -2.806      0.000     -2.806    -2.806
==============================================================
[NOTE]: R-squared in fuzzy regression does not fall between
 0 and 1 because sum-of-square indentity theorem (SSTO =
 SSE + SSR) does not hold
```

(b) HBS

Fig. 5. The fuzzy regression analysis with (a) Tan (b) HBS formulation

Fig. 5 shows two fuzzy linear regression results with Tan formulation and HBS formulation. The value of jaccard similarity, R^2 in Tan formulation is both lower than those of HBS formulation, and also the MAPE of Tan is higher than that of HBS. In this case, we may conclude that HBS formulation outperforms Tan formulation.

III. Case study

To evaluate the performance of the proposed fuzzy linear regression module, we conduct an empirical analysis of forecasting the electricity demand of a manufacturing factory in Shanghai. The standard predictors are production quantities per day, the weather information in each day, such as temperature, precipitation, and wind speed, and the qualitative variable like quarter (Q1, Q2, Q3, and Q4). The response is the electricity usage per day, and the dataset is ranged from April 1st 2019 to December 29th 2019. The dataset is represented as follows:

Location id	**Date**	**Electricity usage (kw/h)**
location_1	2019/4/1	1000
location_1	2019/4/2	1185
…	…	…
location_102	2019/12/29	7
location_102	2019/12/30	8
location_102	2019/12/31	6

(a) Electricity usage for 102 locations

Date	**Production quantity (piece)**
2019/4/2	3250
2019/4/3	1780
…	…
2019/12/27	373146
2019/12/28	122057
2019/12/29	9555

(b) Production quantity

Time	**Temperature (°C)**	**Precipitation (mm)**	**Wind Speed (km/h)**
2019/4/1 00:00	10.72	0	1.553
2019/4/1 01:00	10.33	0	2.716
2019/4/1 02:00	8.32	0	5.068
…	…	…	…
2019/12/31 22:00	3.28	0	2.664
2019/12/31 23:00	2.96	0	1.065

(c) The information of weather

Table. 1. The raw dataset of (a) electricity usage for 102 locations,
(b) production quantity, and (c) the information of weather.

Our goal is to find an appropriate mathematical model of the electricity usage. In order to make the fuzzy linear regression model flexible and convincing, we attempt to merge the conventional linear regression model and fuzzy linear regression into a systematical analysis procedure; therefore, the analytical result may be improved.

IV. Methodology

Conventional regression analysis is a powerful tool to determine the best fitting coefficients of the mathematical model from the dataset, and can offer much more statistical information, including p-value, significance, main effect, interaction effect, confidence region, and others. However, this methodology requires some assumptions of given data, and the effectiveness of its analytical result will be highly affected if the data lacks such assumption like identically normality, constant variance of the error, uncorrelated data. In this case, fuzzy linear regression model can be an alternative. For the sake of the foundation of fuzzy linear regression (that is possibility theory and fuzzy set theorem), fuzzy linear regression methodology does not have those statistical assumptions, and also the unsatisfied assumptions or errors of conventional linear model can be viewed as the fuzziness of the model structure under the fuzzy regression methodology [1]. However, without the support of those statistical assumptions, the fuzzy regression methodology becomes a certain minimization problem, which is hard to conduct the further analysis such as confidence interval, prediction interval, and MSE. As a result, one of the common solutions is to combine fuzzy regression with conventional regression model [5]. In [5], the procedure of analysis is shown as follows:

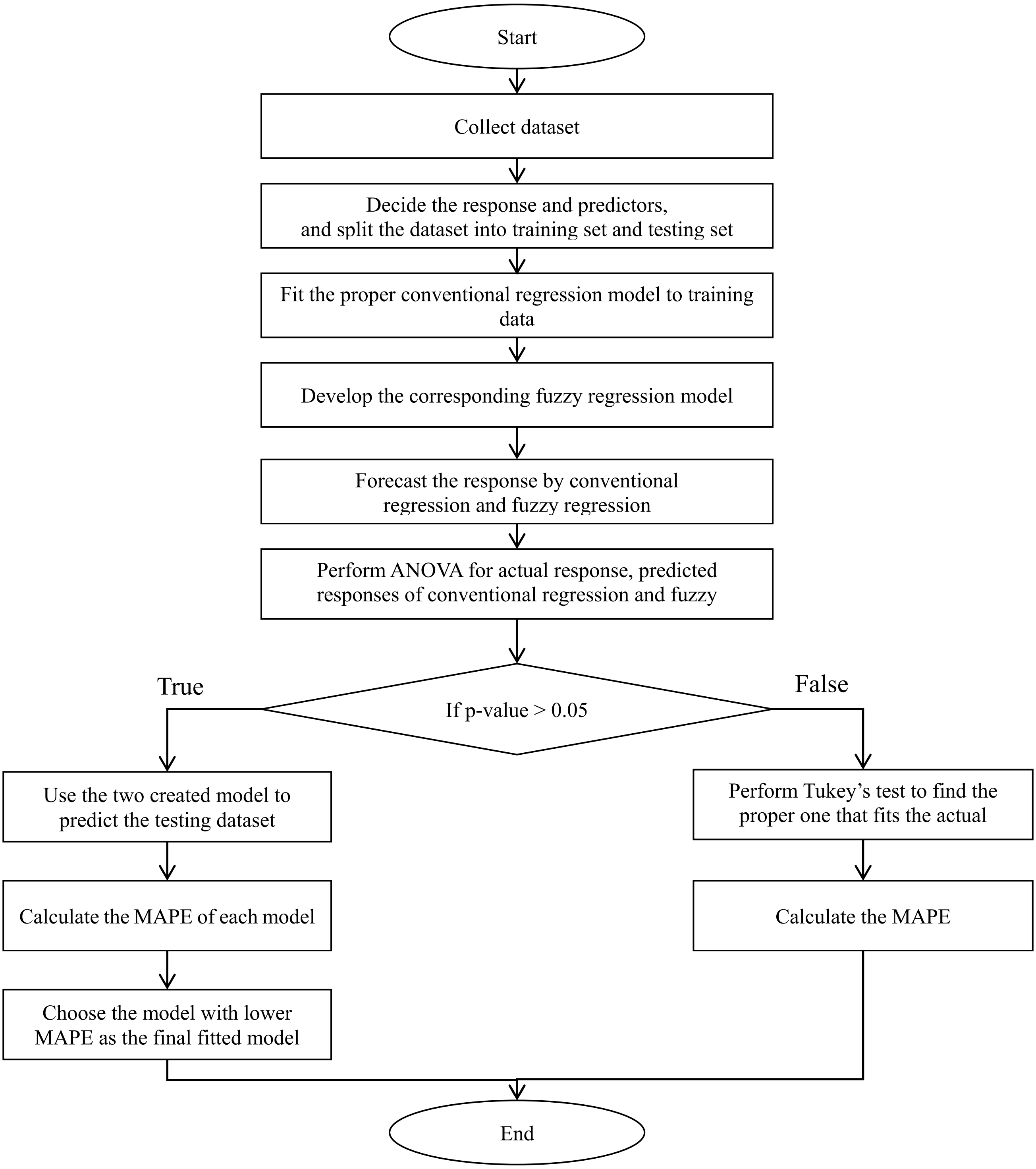

Start
Collect dataset
Decide the response and predictors, and split the dataset into training set and testing set
Fit the proper conventional regression model to training data
Develop the corresponding fuzzy regression model
Forecast the response by conventional regression and fuzzy regression
Perform ANOVA for actual response, predicted responses of conventional regression and fuzzy
If p-value > 0.05
True
False
Use the two created model to predict the testing dataset
Calculate the MAPE of each model
Choose the model with lower MAPE as the final fitted model
Perform Tukey's test to find the proper one that fits the actual
Calculate the MAPE
End

Fig. 6. The procedure of analysis

The first step is to collect the data, decide the proper response and predictors, and separate total dataset into training set and testing set. Due to the advantage of conventional regression methodology that can determine the significant terms by p-value, we apply conventional one first as the foundation of the fuzzy regression methodology. After the fitted conventional regression is introduced, we perform fuzzy regression analysis using the same predictors as that of conventional one. According the two fitted model, we use the predicted value of those two to compare with actual response. The ANOVA analysis is applied to show whether the result has the equal population mean. If null hypothesis is rejected, then we conduct Tukey's test to find out the one that may properly be similar to the actual response, and check the chosen model by testing data with MAPE as the criterion. Otherwise, we use the testing set to validate the performance of those two model and pick up the one with lower MAPE as the final fitted model.

V. Result analysis and performance

In the beginning, we transform the data shown in Table 1 by the following steps. First, pick up the top 20 most electricity consuming locations using median as the criterion. The reason to choose median is that median is robust to outliers and there are some of possible outliers in raw dataset due to the uncertainty in real world. Second, for each location we take the $1.5 \times 99th\ percentile$ as a threshold to filter out unusually large value which is caused by the minor logic error in the electricity usage summary system in the database of company. Third, we sum up the electricity usage of those 20 locations by date as the response for conventional regression and fuzzy regression, and then find out the median of each column in Table 1 (c) per day. In the end, we add a new column 'quarter' called segmentation and merge the pre-processed data into a new dataset like Table 2:

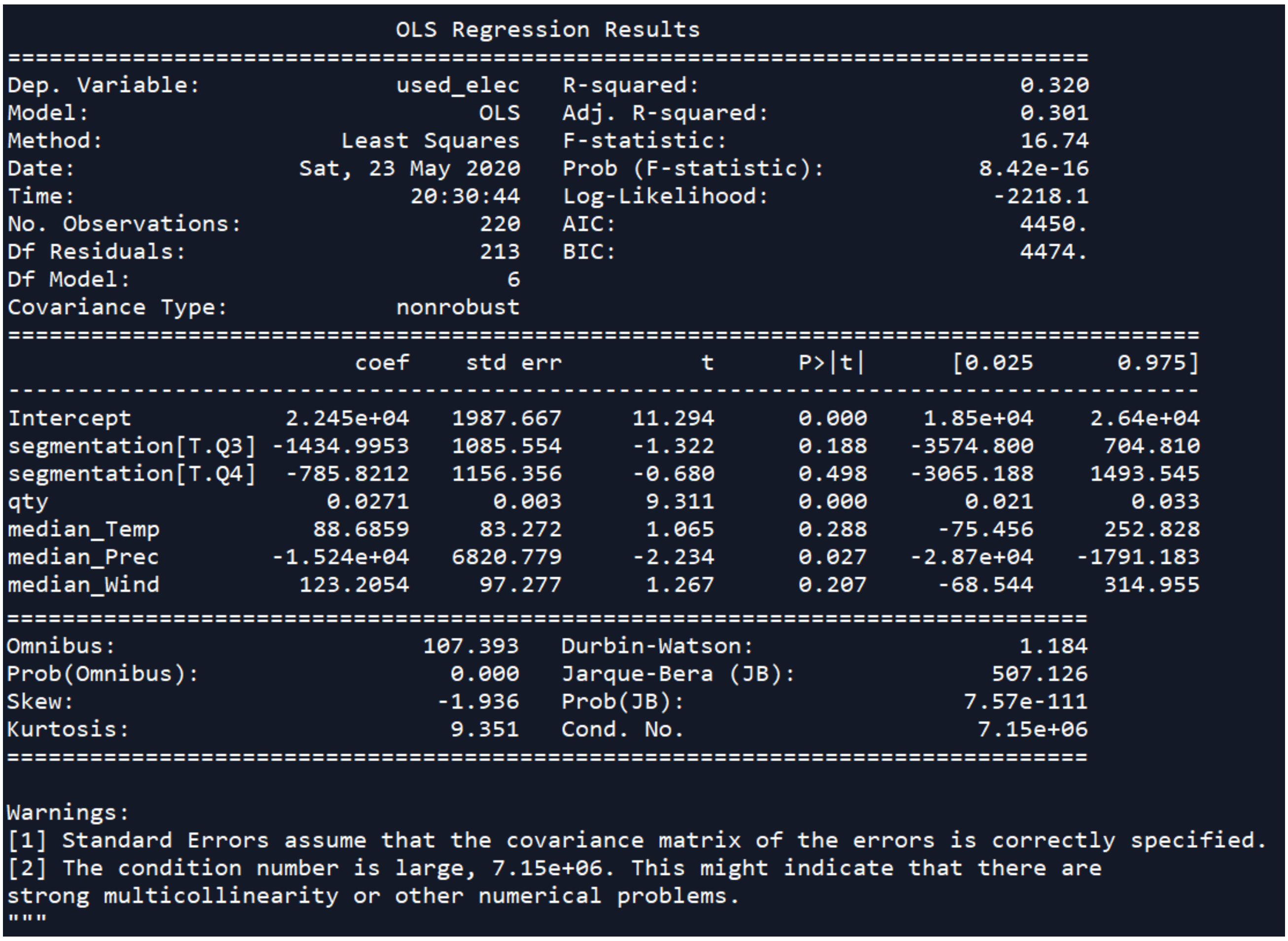
Date	**Electricity usage**	**Segmentation**	**Production quantity**	$\mathbf{Temp_{median}}$	$\mathbf{Prec_{median}}$	$\mathbf{Wind_{median}}$
2019/4/1	36825	Q2	438881	13.0605	0	3.4955
2019/4/2	40362	Q2	3250	10.5255	0	3.6813
…	…	…	…	…	…	…
2019/12/27	38431	Q4	373146	4.3205	0	7.334
2019/12/28	36041	Q4	122057	7.5355	0	9.5507
2019/12/29	24628	Q4	9555	4.7255	0	9.3893

Table. 2. The data matrix of regression methodology.

For simplicity, we assume there is no assumption violation in the full fitted model:

$$Y \sim segmentation + Production\ quantity + Temp_{median} + Prec_{median} + Wind_{median}, \quad \textbf{(16)}$$

and the result is shown in Fig. 7 and 8.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              used_elec   R-squared:                       0.320
Model:                            OLS   Adj. R-squared:                  0.301
Method:                 Least Squares   F-statistic:                     16.74
Date:                Sat, 23 May 2020   Prob (F-statistic):           8.42e-16
Time:                        20:30:44   Log-Likelihood:                -2218.1
No. Observations:                 220   AIC:                             4450.
Df Residuals:                     213   BIC:                             4474.
Df Model:                           6
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept           2.245e+04   1987.667     11.294      0.000    1.85e+04    2.64e+04
segmentation[T.Q3] -1434.9953   1085.554     -1.322      0.188   -3574.800     704.810
segmentation[T.Q4]  -785.8212   1156.356     -0.680      0.498   -3065.188    1493.545
qty                    0.0271      0.003      9.311      0.000       0.021       0.033
median_Temp           88.6859     83.272      1.065      0.288     -75.456     252.828
median_Prec        -1.524e+04   6820.779     -2.234      0.027    -2.87e+04   -1791.183
median_Wind          123.2054     97.277      1.267      0.207     -68.544     314.955
==============================================================================
Omnibus:                      107.393   Durbin-Watson:                   1.184
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              507.126
Skew:                          -1.936   Prob(JB):                    7.57e-111
Kurtosis:                       9.351   Cond. No.                     7.15e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.15e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Fig. 7. The analysis result of OLS

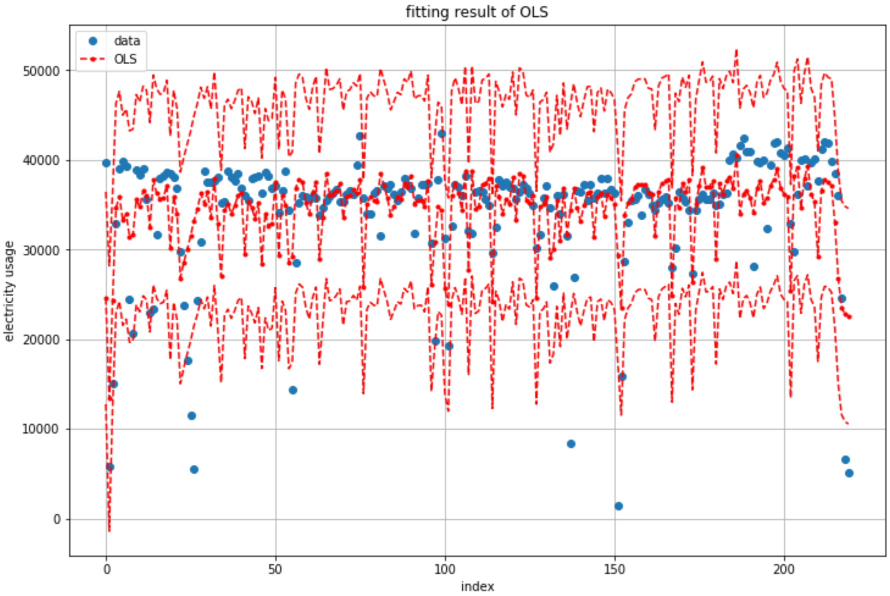
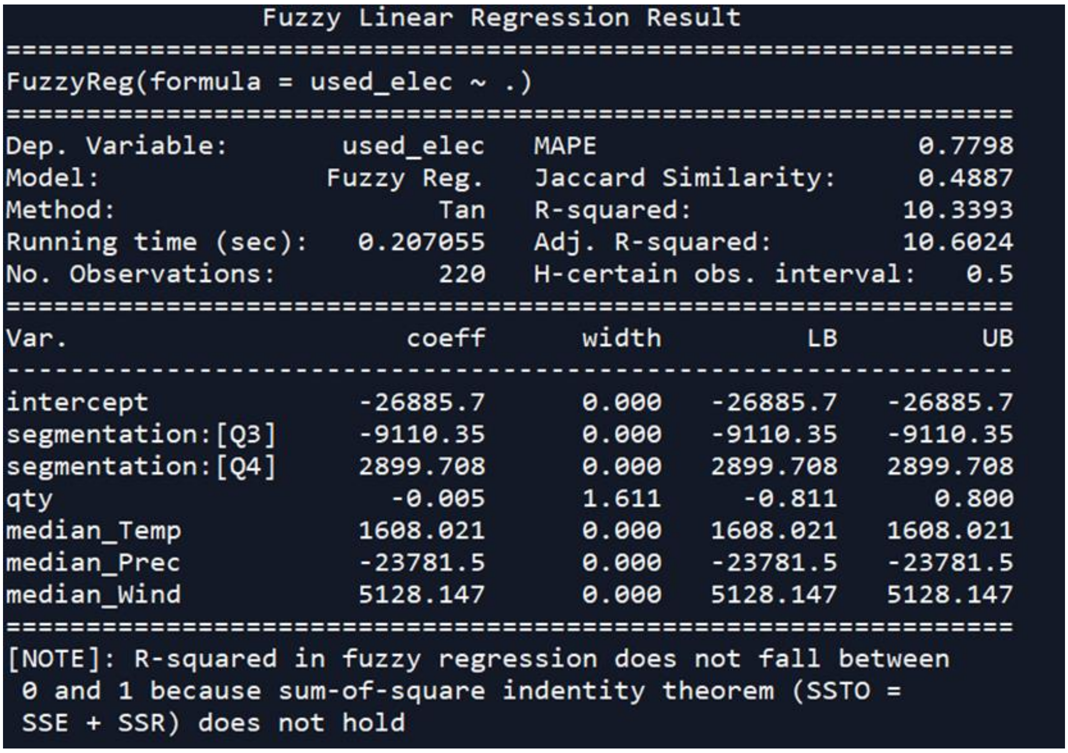


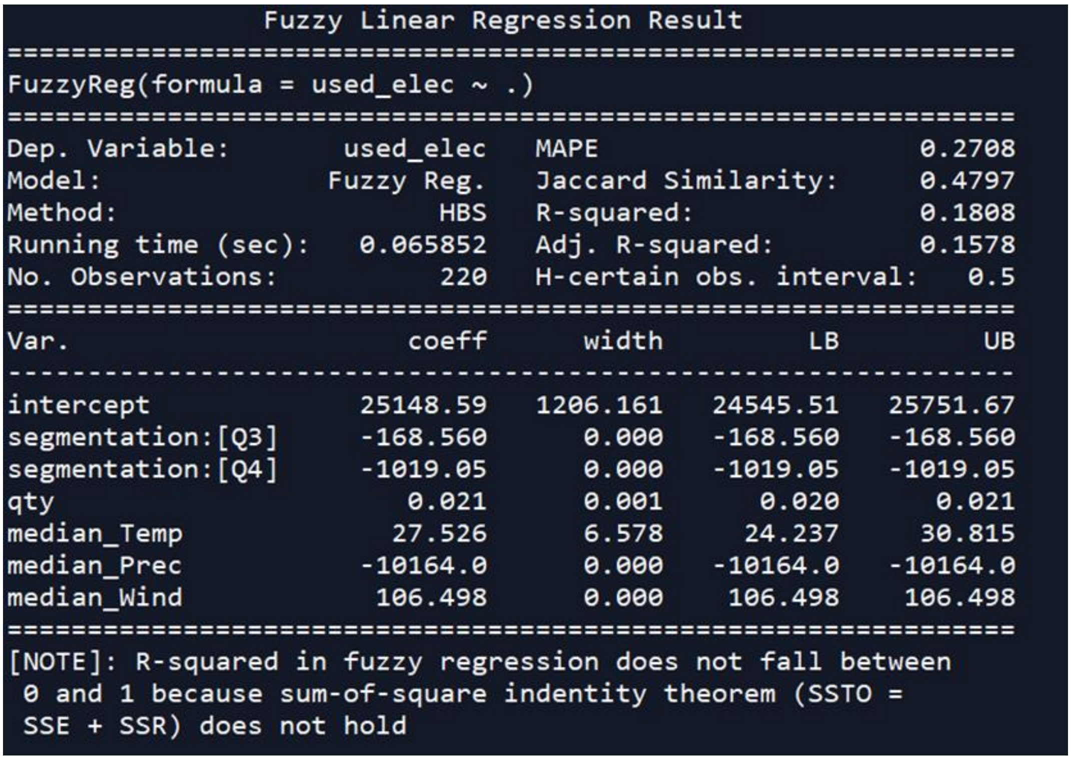
Fig. 8. The visualization of fitting result of OLS

According to model (16), we apply fuzzy regression to fit the same model with Tan and HBS formulation and assume that response y_i is a fuzzy triangular number with the width of $\frac{y_i}{20}$, $i = 1,\ldots,220$. Note that the following fuzzy regression models have added the corresponding constraint mentioned in equation (11). (see the Appendix B to check the fuzzy regression result with additional constraint.)



```
                Fuzzy Linear Regression Result
================================================================
FuzzyReg(formula = used_elec ~ .)
================================================================
Dep. Variable:        used_elec   MAPE                    0.7798
Model:               Fuzzy Reg.   Jaccard Similarity:     0.4887
Method:                     Tan   R-squared:             10.3393
Running time (sec):    0.207055   Adj. R-squared:        10.6024
No. Observations:           220   H-certain obs. interval:   0.5
================================================================
Var.                      coeff      width        LB          UB
----------------------------------------------------------------
intercept              -26885.7      0.000  -26885.7    -26885.7
segmentation:[Q3]      -9110.35      0.000  -9110.35    -9110.35
segmentation:[Q4]      2899.708      0.000  2899.708    2899.708
qty                      -0.005      1.611    -0.811       0.800
median_Temp            1608.021      0.000  1608.021    1608.021
median_Prec            -23781.5      0.000  -23781.5    -23781.5
median_Wind            5128.147      0.000  5128.147    5128.147
================================================================
[NOTE]: R-squared in fuzzy regression does not fall between
 0 and 1 because sum-of-square indentity theorem (SSTO =
 SSE + SSR) does not hold
```

(a) Tan



```
                Fuzzy Linear Regression Result
================================================================
FuzzyReg(formula = used_elec ~ .)
================================================================
Dep. Variable:        used_elec   MAPE                    0.2708
Model:               Fuzzy Reg.   Jaccard Similarity:     0.4797
Method:                     HBS   R-squared:              0.1808
Running time (sec):    0.065852   Adj. R-squared:         0.1578
No. Observations:           220   H-certain obs. interval:   0.5
================================================================
Var.                      coeff      width        LB          UB
----------------------------------------------------------------
intercept              25148.59   1206.161  24545.51    25751.67
segmentation:[Q3]      -168.560      0.000  -168.560    -168.560
segmentation:[Q4]      -1019.05      0.000  -1019.05    -1019.05
qty                       0.021      0.001     0.020       0.021
median_Temp              27.526      6.578    24.237      30.815
median_Prec            -10164.0      0.000  -10164.0    -10164.0
median_Wind             106.498      0.000   106.498     106.498
================================================================
[NOTE]: R-squared in fuzzy regression does not fall between
 0 and 1 because sum-of-square indentity theorem (SSTO =
 SSE + SSR) does not hold
```

(b) HBS

Fig. 9. The analysis result of fuzzy regression (a) Tan (b) HBS

In Fig. 9, it shows the numerical results and the summary table is represented in Table 3.

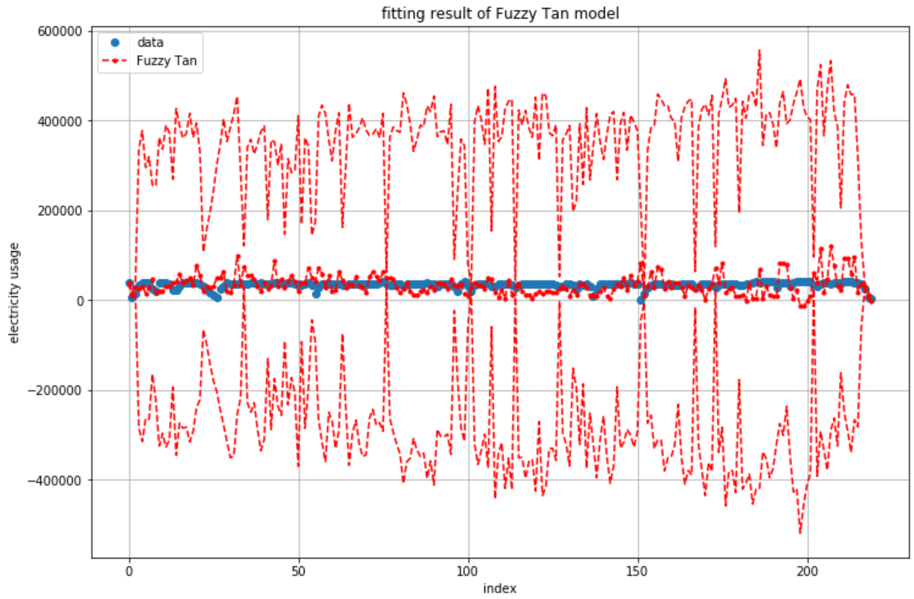
		OLS	**Tan**	**HBS**
criterion	R-squared	0.32	10.34	0.1808
	Adj. R-squared	0.301	10.6	0.1578
	MAPE	0.2646	0.78	0.2708

coefficient	intercept	22450	-26886	25148
	segmentation[Q3]	-1435	-9110	-168.56
	segmentation[Q4]	-785	2900	-1019.05
	production quantity	0.0271	-0.05	0.021
	median_Temp	88.68	1608.021	27.526
	median_Prec	-15240	-23782	-10164
	meidan_Wind	123.2	5128	103.498

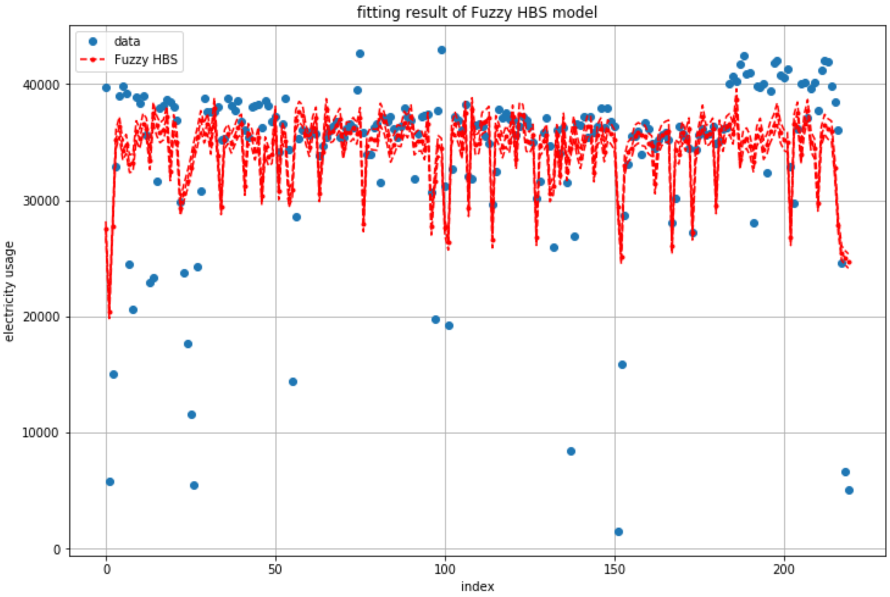
Table. 3. The summary table of 3 models.

As mentioned that R-squared does not follow between 0 and 1 of fuzzy regression model; therefore, the R-squared shown here is just one of the available criterion for validation. Notice that the R-squared of Tan is greater than 1, which implies that SSR is higher than $SSTO$ (see equation (12)). As a result, we may conclude that the fitting result of Tan is worse than the fitting result of the mean value of the electricity usage. In the fact that Tan formulation is sensitive to outliers, the data matrix does contain some outliers; therefore, the fitting result of Tan model is not ideal (see Fig. 10 (a)). For the following discussion we exclude the Tan model due to its inferior performance.

The R-squared, adjusted R-squared of OLS both are higher than those of HBS model; however, the MAPE of OLS does not outperform that of HBS. In consequence, we may not conclude which model is better and Fig. 11 indicates that OLS is quite similar to HBS model.



(a) Tan



(b) HBS

Fig. 10. The visualization of fitting result of fuzzy regression (a) Tan (b) HBS.

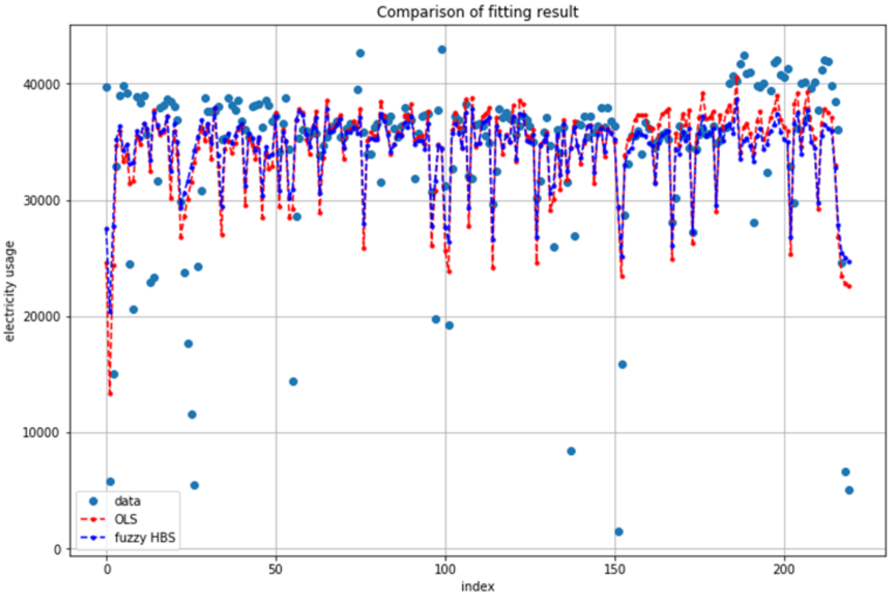
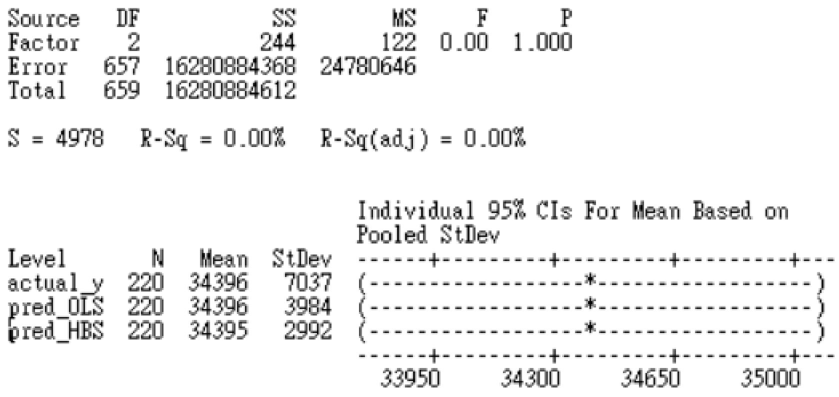


Fig. 11. The comparison of fitting result.

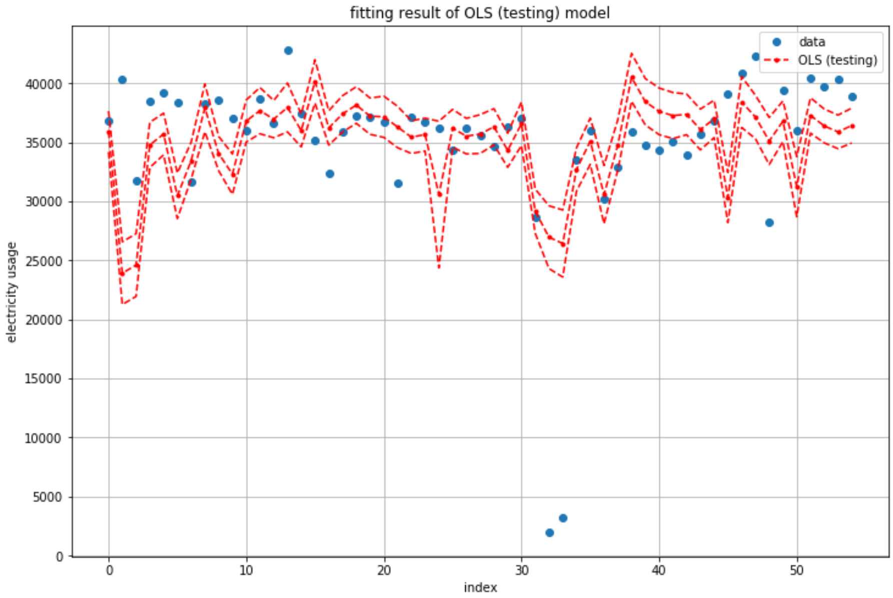
In order to determine the better model between these two, ANOVA is introduced to check the population mean between the actual response and predicted responses of OLS and HBS, shown in Fig. 12:



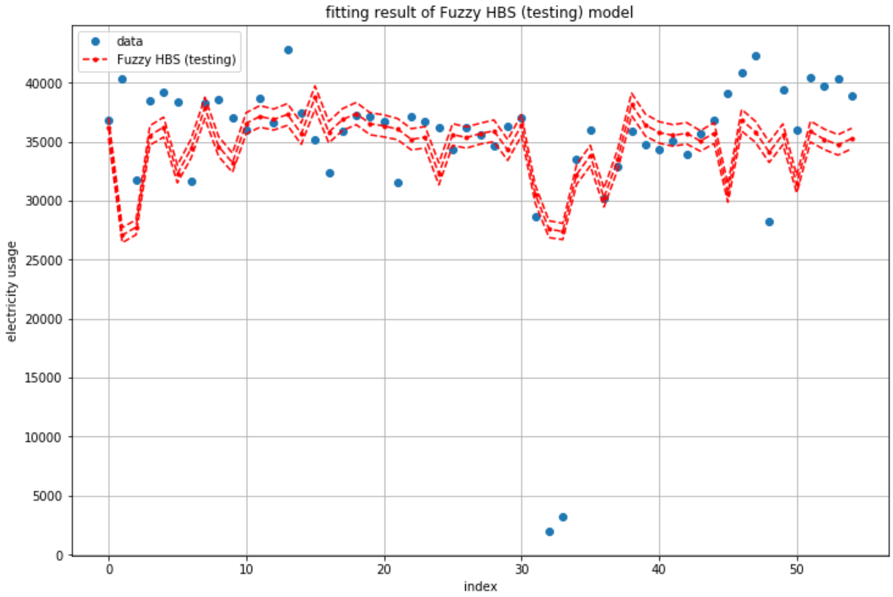
```
Source    DF           SS        MS     F      P
Factor     2          244       122  0.00  1.000
Error    657  16280884368  24780646
Total    659  16280884612

S = 4978   R-Sq = 0.00%   R-Sq(adj) = 0.00%

                               Individual 95% CIs For Mean Based on
                               Pooled StDev
Level       N   Mean  StDev  ------+---------+---------+---------+---
actual_y  220  34396   7037  (-----------------*-----------------)
pred_OLS  220  34396   3984  (-----------------*-----------------)
pred_HBS  220  34395   2992  (-----------------*-----------------)
                             ------+---------+---------+---------+---
                               33950     34300     34650     35000
```

Fig. 12. The result of ANOVA.

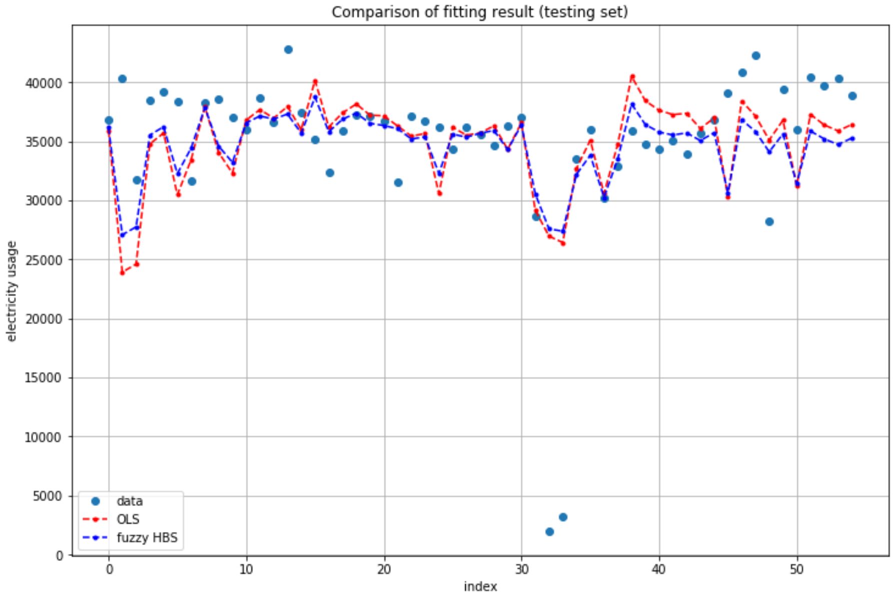
Result in Fig. 12 does not reject the null hypothesis, which indicates that 3 types of values have the same population mean. Thus, we use testing set to validate the models. The MAPE of OLS and HBS are 0.437 and 0.441 respectively, and their difference is 0.004, which is a quite small value. We may conclude that these two models have the same performance (see Fig. 13).



(a) OLS



(b) HBS



(c) Comparison

Fig. 13. The visualization of fitting result of testing set (a) OLS (b) HBS (c) Comparison.

VI. Conclusion

In this project, we have implemented a fuzzy linear regression module and conducted a case study of forecasting electricity demand with the proposed module and conventional linear regression methodology. In the implementation process, we have created 2 kinds of fuzzy regression methodology with 2 different LP formulations, and introduced the sum coding to deal with qualitative variables, which allows the module to solve the model that contains any dummy variable. Moreover, the module is put in an additional constraint described in equation (11), which can improve the fitting result.

We have tried the fuzzy regression module on a case study. The analysis result has shown that fuzzy linear regression with HBS formulation is as good as the conventional linear regression methodology. Also, there is no any statistical assumption in fuzzy linear regression; therefore, it is no need to do the residual analysis like OLS must do. As a result, it is more convenient for researchers to take advantage of. However, there are some limitations in the fuzzy linear regression module. First, the response and coefficients of fuzzy regression model are supposed to be symmetric triangular fuzzy numbers. If other kinds of fuzzy number are included in the model, it requires a more complicated treatment or solution. Second, in order to choose the proper predictors in the fuzzy linear regression, it relies on the model of OLS in advance. Without the help of OLS model, the fuzzy linear regression has no ability to determine the proper predictors.

REFERENCES

[1] Tanaka, S., Asai, H. T. S. U. K., & Uegima, K. (1982). Linear regression analysis with fuzzy model. IEEE Trans. Systems Man Cybern, 12, 903-907.

[2] Hojati, M., Bector, C. R., & Smimou, K. (2005). A simple method for computation of fuzzy linear regression. European Journal of Operational Research, 166(1), 172-184.

[3] Sarkar, M. R., Rabbani, M. G., Khan, A. R., & Hossain, M. M. (2015, May). Electricity demand forecasting of Rajshahi City in Bangladesh using fuzzy linear regression model. In 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (pp. 1-3). IEEE.

[4] Gorgulu, O., Akilli, A., Gorgulu, O., Gorgulu, O., & Akilli, A. (2018). Estimation of 305-days milk yield using fuzzy linear regression in Jersey dairy cattle. JAPS, Journal of Animal and Plant Sciences, 28(4), 1174-1181.

[5] Azadeh, A., Khakestani, M., & Saberi, M. (2009). A flexible fuzzy regression algorithm for forecasting oil consumption estimation. Energy Policy, 37(12), 5567-5579.

Appendix A

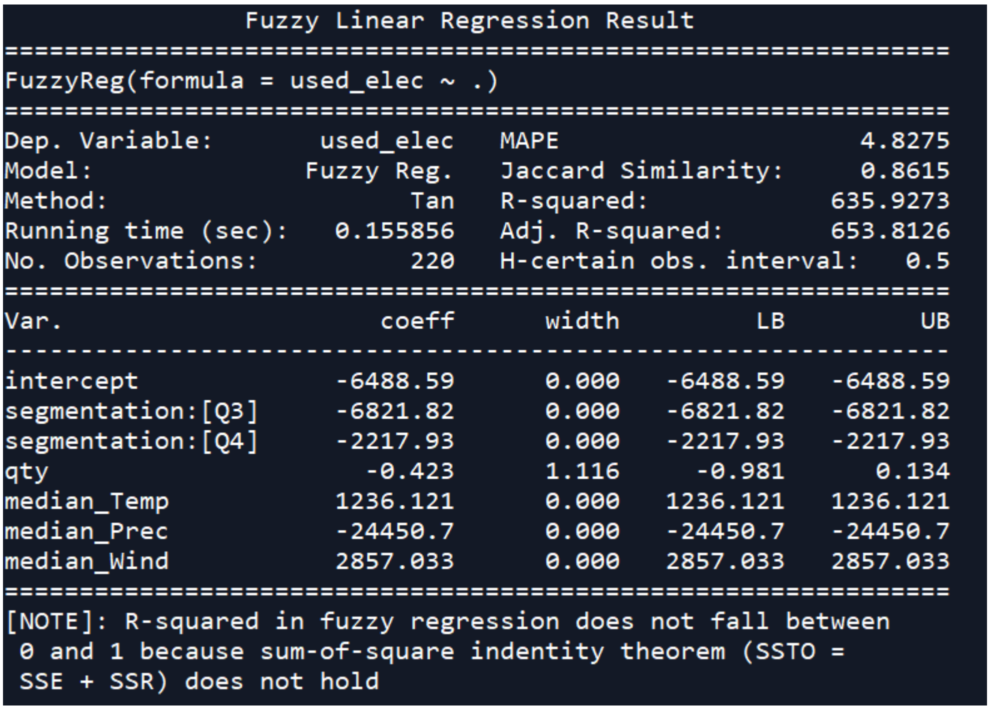
The proof that the R^2 of the fuzzy regression does not hold the theorem of Sum-of-Square Identity:

$$
\begin{aligned}
SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}\right)^2 \\
&= \sum_{i=1}^{n}\left[\left(Y_i - \hat{Y}_i\right)^2 + 2\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) + \left(\hat{Y}_i - \bar{Y}\right)^2\right] \\
&= \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + 2\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 \\
&= \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + 2\sum_{i=1}^{n} e_i\left(\hat{Y}_i - \bar{Y}\right) + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 \\
&= \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + 2\sum_{i=1}^{n} e_i\hat{Y}_i - 2\bar{Y}\sum_{i=1}^{n} e_i + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 \\
&= SSE + 2\sum_{i=1}^{n} e_i\hat{Y}_i - 2\bar{Y}\sum_{i=1}^{n} e_i + SSR,
\end{aligned}
$$

Two terms in red are not necessary to be zero; therefore, $SSTO = SSE + SSR$ is not always hold.

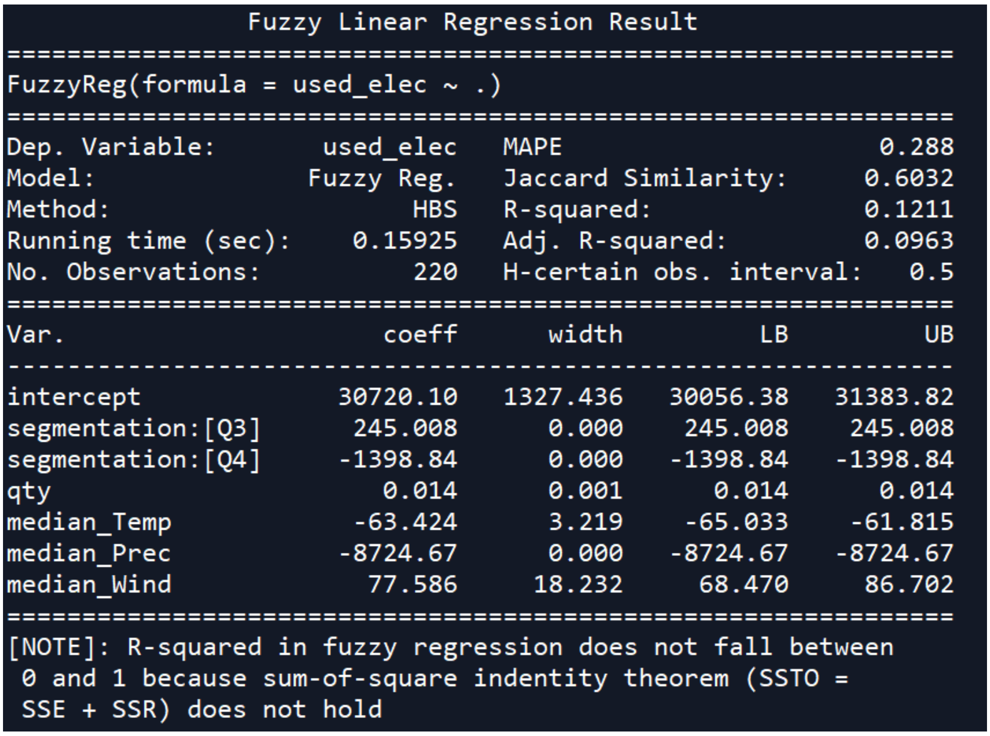
Appendix B

The fuzzy regression model without additional constraint, that is, equation (11):



```
                Fuzzy Linear Regression Result
===============================================================
FuzzyReg(formula = used_elec ~ .)
===============================================================
Dep. Variable:         used_elec   MAPE                   4.8275
Model:                Fuzzy Reg.   Jaccard Similarity:    0.8615
Method:                      Tan   R-squared:           635.9273
Running time (sec):     0.155856   Adj. R-squared:      653.8126
No. Observations:            220   H-certain obs. interval:  0.5
===============================================================
Var.                       coeff       width        LB        UB
---------------------------------------------------------------
intercept               -6488.59       0.000  -6488.59  -6488.59
segmentation:[Q3]       -6821.82       0.000  -6821.82  -6821.82
segmentation:[Q4]       -2217.93       0.000  -2217.93  -2217.93
qty                       -0.423       1.116    -0.981     0.134
median_Temp             1236.121       0.000  1236.121  1236.121
median_Prec             -24450.7       0.000  -24450.7  -24450.7
median_Wind             2857.033       0.000  2857.033  2857.033
===============================================================
[NOTE]: R-squared in fuzzy regression does not fall between
 0 and 1 because sum-of-square indentity theorem (SSTO =
 SSE + SSR) does not hold
```

(a) Tan



```
                Fuzzy Linear Regression Result
===============================================================
FuzzyReg(formula = used_elec ~ .)
===============================================================
Dep. Variable:        used_elec   MAPE                     0.288
Model:               Fuzzy Reg.   Jaccard Similarity:     0.6032
Method:                     HBS   R-squared:              0.1211
Running time (sec):     0.15925   Adj. R-squared:         0.0963
No. Observations:           220   H-certain obs. interval:   0.5
===============================================================
Var.                      coeff      width         LB         UB
---------------------------------------------------------------
intercept              30720.10   1327.436   30056.38   31383.82
segmentation:[Q3]       245.008      0.000    245.008    245.008
segmentation:[Q4]      -1398.84      0.000   -1398.84   -1398.84
qty                       0.014      0.001      0.014      0.014
median_Temp             -63.424      3.219    -65.033    -61.815
median_Prec            -8724.67      0.000   -8724.67   -8724.67
median_Wind              77.586     18.232     68.470     86.702
===============================================================
[NOTE]: R-squared in fuzzy regression does not fall between
 0 and 1 because sum-of-square indentity theorem (SSTO =
 SSE + SSR) does not hold
```

(b) HBS

Appendix. B. The analysis result of fuzzy regression without an additional constraint (a) Tan (b) HBS

	Tan			HBS		
Add an additional constraint	R-squared	Adj. R-squared	MAPE	R-squared	Adj. R-squared	MAPE
True	10.340	10.600	0.780	0.181	0.158	0.271
False	635.927	653.813	4.828	0.121	0.096	0.288

Appendix. B. The summary table of fuzzy regression with/without an additional constraint

The summary table above shows that the model with an additional constraint is better because MAPE of model with an extra constraint is lower than that of model without an extra constraint. Moreover, the R-squared, and adjusted R-squared of Tan model with/without a constraint are both higher than 1. In this case, the model with smaller R-squared and adjusted R-squared is better, which indicates that the model with an additional constraint is better. On the other hands, HBS models with/without an extra constraint have R-squared, and adjusted R-squared, lower than 1. In this situation, the model with higher R-squared and adjusted R-squared is better, which implies that the model with an additional constraint is better.

Furthermore, we use the testing set to evaluate these two cases. The performances of Tan model with/without an extra constraint are terrible; therefore, we only consider the HBS model.

	HBS
Add an additional constraint	MAPE
True	0.441

False	0.472

Appendix. B. The summary table of HBS model with/without an additional constraint on testing set

According to MAPE, the summary table shows that HBS model with an additional constraint outperforms the one without an extra constraint.