國立政治大學應用數學系

碩士學位論文

模糊數據的局部加權回歸

Locally Weighted Regression of Fuzzy Data

碩士班學生：陳 帥　撰

指導教授：吳柏林 博士

中華民國 106 年 06 月 12 日

致 謝

短短的兩年時間，在許多人的幫助之下，即將完成碩士班學位論文的撰寫。回想起來，這段經歷竟然是如此的特殊。向在此過程中給予我無私幫助的各位老師表示感謝。尤其是吳柏林老師，在每次的課程之中，都給予我充分的、專業的指導與幫助。

此外，對在口試過程中給予我建議與意見的陳瑞照、曾正男兩位委員致以謝意；以及，對在日常學習生活中給予我幫助的各位應數系同學表示感謝與祝福，謝謝你們。

陳帥 謹致于

國立政治大學應用數學系 碩士班

中華民國 106 年 6 月

摘要

目標：本文旨在建構一種新型的模糊回歸模式，解決一类較複雜的模糊回歸問題。

研究方法：推廣局部加權回歸的思想，先從理論上構建新模型；然後借由模拟數據，從多個方面考察新模型的性質，并和其他模型做比較。

發現：局部加權回歸方法結合模糊隸屬度概念，使模糊回歸理論有更多的應用場合。

原創性：目前在模糊回歸領域的主流思想是通過線性規劃等方法來構建模型，而本文另闢蹊徑，首次從局部加權的角度構建了模糊回歸的新模型。

關鍵字：模糊理論 模糊回歸分析 局部加權

Abstract

Objective: This paper aims to construct a new fuzzy regression model to solve a more complex fuzzy regression problem.

Method: Build a new model by promoting the idea of locally weighted regression; Using simulated data to compare the new model with other models.

Conclusion: The fuzzy membership degree concept combined with the locally weighted regression method makes the fuzzy regression theory have more applications.

Originality: At present, the main idea in the field of fuzzy regression is to construct models by means of linear programming. In this paper, a new model of fuzzy regression is constructed from the perspective of locally weighted method for the first time.

Keyword: Fuzzy theory、 Fuzzy regression、Locally weighted method

目錄

1. 前言

模糊理論最早由美國加州大學伯克萊分校教授 L.A.Zadeh 提出，旨在運用模糊集合來更好地描述處理現實環境中的各種不確定（uncertainty）與模糊性（fuzziness）資料[1]。給定恰當的隸屬度函數后可以定義模糊數的概念。常見的模糊數有三角模糊數（Triangular Fuzzy Number, TFN），及梯形模糊數（Trapezoidal Fuzzy Number，TrFN）等。

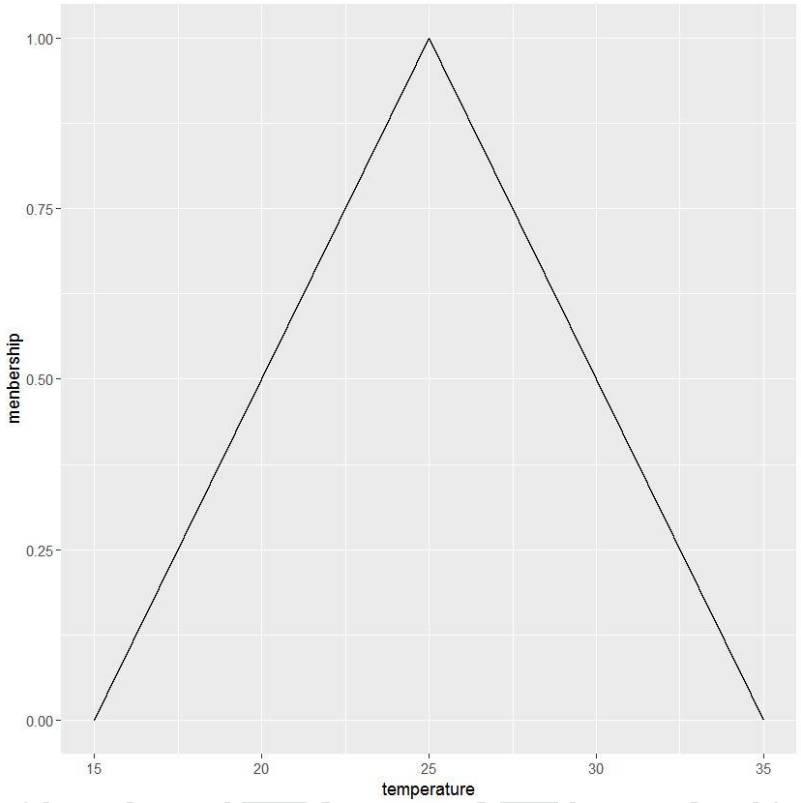
一個簡單的實例如下：當某人被問及“使你感覺舒適的溫度是多少”時，如果限定只能用一個實數來回答，或許他給出的答案是 25℃。然而，當溫度是 20℃時，該人也並未感覺不適。所以单個實數不能很好地描述該人的想法。或許，採用模糊數來回答更能反映實際情況:令該人感到舒適的溫度是個三角模糊數，25℃是該三角模糊數的頂點，即“25℃”這個點屬於“令該人感覺舒適的溫度”這個集合的隸屬程度達到最高的 1；當溫度慢慢降低到 15℃或慢慢增高到 35℃時，該隸屬度直線遞減為 0，也就是說“20℃”屬於“令人感覺舒適的溫度”這個集合的隸屬度為 0.5。這樣的描述或許更符合該人的真實想法。三角型模糊數的圖示見圖 1。

模糊理論目前已經被廣泛地運用在統計學的各個領域，例如模糊決策分析（fuzzy decision making analysis）方面有 Zhi-Ping Fan(2002)，Zeshui Xu（2008）等；模糊統計聚類（fuzzy clustering）方面有 KL Wu & MS Yang (2005)， Keh-Shih Chuang(2006)，Shihua Zhang（2007）等；模糊回歸分析方面有 Reshma Khemchandani（2009），M. Hadi Mashinchi（2011）等；模糊時間序列（fuzzy time series）方面有 Kunhuang Huarng（2006），Mehdi Khashei（2008）等；這些研究都取得了不錯的成果。

在回歸分析領域，傳統的回歸模式裡自變數與因變數皆是實數，觀察值的不確定性來自隨機現象；然而，如果考慮觀察值的不確定性是來自多重隸屬現象，也就是說因變數是一個帶有隸屬度信息的模糊數，那麼傳統針對實數數據的回歸模式就需要修正了。運用模糊理論，套用回歸分析的方法來處理資料模糊的問題，就叫做模糊回歸分析。早在 1982 年，H. Tanaka, S. Uejima 和 K. Asai 就提出

了模糊回歸的概念[2]。目前，構建模糊回歸模式的方法主要有兩種：線性規劃法和最小平方法。比起線性規劃法，最小平方法較能符合誤差隨機分佈的精神[3]。

圖 1：一個三角型模糊數示意圖



而本文旨在構建一種新的模糊回歸模式，它和最小平方法有緊密聯繫但又有所不同。

模糊線性回歸常表示成：

$$y = \beta_0 + \beta_1 x \qquad ①$$

其中 x 是自變數，y 是因變數。如果給出 n 組樣本$(x_i, y_i); i = 1, \dots, n$，我們要做的就是去得到參數$\beta = (\beta_0, \beta_1)^T$的估計。在本文中，我們主要關注自變數是傳統實數數據，因變數是模糊數據的情況。

在之後的模型構建中，將認定模糊因變數 y 是區間模糊數，並把它表示成“中心點+半徑”的形式。這是為了能更清晰地介紹新模型。事實上，無論是三角型模糊數，梯形模糊數，都可以採用相同的“中心點+半徑”的表示方法，見例 1。所以本文介紹的方法並不局限於區間模糊數的情況。

例 1：如圖 1 所示是一個三角型模糊數，按照“中心點+半徑”的思路，可以

將其表示成<25；-10,0,10>。即中心點是 25，中心點加上三個半徑的值就可以得到三個端點的值。中心點可由各個端點的簡單算術平均得到。顯然對於梯形模糊數也可以做如此改寫[4]。理由如下：

（1） 作者認為中心點決定了該模糊數“大體處於什麼位置”，是模糊數的“位置”參數；而半徑蘊含了模糊數的隸屬度信息，決定了該模糊數隸屬度的值，是模糊數的“尺度”參數。尤其是在有不同側重點的情況下，中心點和半徑應該分開來研究。

（2） “中心點+半徑”的表示形式能夠很好地描述一類模糊數，沒有造成信息遺漏。

穩健局部加權回歸(Robust Locally Weighted Regression)由美國著名統計學家、電腦科學教授 Willian S. Cleveland 與 1979 年提出，是一種基於最小平方思想，能夠在實數數據密集的情況下表現良好的方法。它的主要思想仍是基於“最小平方”的原則，如下所示：

首先記 W 為滿足如下條件的某個函數，并稱之為權重函數：

1. $W(x) > 0$ for $|x| < 1$;
2. $W(-x) = W(x)$;
3. W(x)在 x 大於等於 0 時非增；
4. $W(x) = 0$ for $|x| \geq 1$;

按一定要求選擇$0 < f \leq 1$，令 r 為最接近n·f的整數。對每一個自變數能取值的點 x 都能定義一系列的權重$W_x(x_i) = W(x_i - x)$。這裡的權重函數經過按比例放縮，放縮后的權重函數滿足：當x_i恰好是離 x 第 r 近的點時，$W_x(x_i)$第一次為 0。另外，為了使模型更有穩健性，減少極端值對結果的影響，按一定標準構建係數δ_i，這是一個隨著$|y_i - \hat{y_i}|$的增大而減小的值，也即極端值對模新的影響被係數δ_i控制住了，這也保證了該模型具有穩健（robust）的性質。

假設$\beta_x = (\beta_{x0}, \beta_{x1}, \dots, \beta_{xp})^T$為待估參數，那麼它的估計就為

$$\widehat{\beta_x} = \arg\min_{\beta_x} \{\sum_{i=1}^{n} \delta_i W_x(x_i)(y_i - \beta_x^T X_i)^2\}$$

如此，我們得到了在點 x 處對因變量的估計$y_x = \widehat{\beta_x} \cdot x$。在每個點 x 處，都按

照之上的程序，就可以得到一系列的因變量的估計，進而得到一條光滑的曲線，這就是穩健局部加權回歸的基本理論過程[5]。

值得一提的是，該方法實際上是基於某種回歸模型而做的改進（比如基於普通線性回歸）；在後文採用該方法思想來構建模糊情況下的回歸新模型的時候，也是基於線性模糊回歸模型①的，後面不再次敘述。

2. 模糊數據的局部加權回歸

2.1 模型的建構

給出 n 個樣本為$(x_i, y_i); \mathrm{i} = 1, \ldots, \mathrm{n}$，其中$x_i$是實數類型的自變數數值，而$y_i =< c_i, r_i >$為模糊數應變數。給出模型如下：

$$< c_i, r_i >=< \beta_{c0}, \beta_{r0} > + x_i \cdot < \beta_{c1}, \beta_{r1} >, i = 1, \ldots, n$$

模型里的模糊數都是“中心點+半徑”的表示形式，待估參數為：

$$< \hat{\beta}_c, \hat{\beta}_r >=< (\hat{\beta}_{c0}, \hat{\beta}_{c1})^T, (\hat{\beta}_{r0}, \hat{\beta}_{r1})^T >$$

也就是說，待估參數也是個模糊數，並且它的中心點與半徑都是個實數向量。

除此之外，有必要給出實數與模糊數做乘法的定義。

定義 1 令a為一實數，$< \mathrm{c}, \mathrm{r} >$為以“中心點+半徑”表示的模糊數，那麼該實數與模糊數的乘積定義為：

$$< c', r' >=< \mathrm{a} \cdot \mathrm{c}, \mathrm{a} \cdot \mathrm{r} >= \mathrm{a} \cdot < c, r >$$

我們先關注給出的區間模糊數的 n 個中心點，記為$c_i; i = 1, \ldots, n$。此時，給定$x_0 \in [\min\{x_1, \ldots, x_n\}, \max\{x_1, \ldots, x_n\}]$，令$\Delta \mathrm{x} = (|x_1 - x_0|, \ldots, |x_n - x_0|)^T$，記 M 為向量Δx所有元素里的最大值，對自變數 x 以及Δx做如下放縮：

$$\Delta x' = \Delta x \cdot \frac{10}{M}; x_0' = x_0 \cdot \frac{10}{M}$$

那麼就可以在點x_0定義對每個x_i的n個權重為：

$$W_{x_0}(x_i) = W(\Delta x_i')$$

其中 $$\mathrm{W}(\mathrm{x}) = \exp\{-x^2/2\}$$

進行放縮的原因是為了保證每個$\Delta x_i'$的值都在 0-10 之間，注意到：

$$\mathrm{W}(10) \approx 0; \mathrm{W}(0) \approx 1$$

圖 2 是上述的權重函數在 0 到 10 的取值情況，可以看出當進行如上步驟的放縮后，與選定點x_0的距離大約為 3 之後的點的權重近似為 0，這樣的放縮也可以使得這n個權重大小充分拉開距離，能夠體現出“權重”的意義。至於放縮程度的大小，實際上可以通過在權重函數的指數部分加參數來控制，如下：

$$W(x) = \exp\{-x^2/(2\gamma^2)\}$$

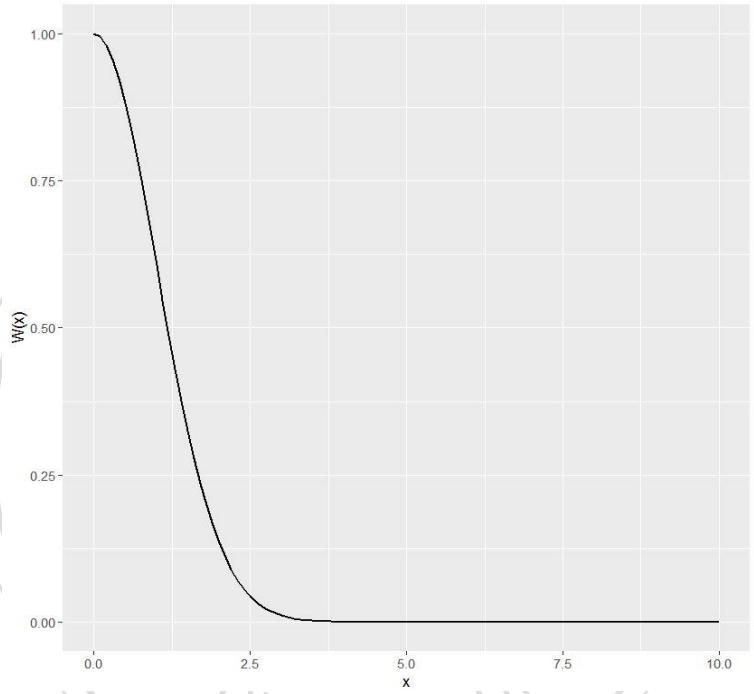
所以這個放縮大小的程度可以視具體情況來調整參數。

從而我們就可以得到在點x_0處對相應區間因變數的中心點c_{x_0}的預測：

$$\widehat{c_{x_0}} = \widehat{\beta_{x_0}}_0 + \widehat{\beta_{x_0}}_1 \cdot x_0'$$

其中$\widehat{\beta_{x_0}} = (\widehat{\beta_{x_0}}_0, \widehat{\beta_{x_0}}_1)^T = \underset{(\widehat{\beta_{x_0}}_0, \widehat{\beta_{x_0}}_1)}{\arg min} \left\{\sum_{i=1}^{n} W_{x_0}(x_i) \cdot \left(c_i - \widehat{\beta_{x_0}}_0 - \widehat{\beta_{x_0}}_1 \cdot x_i\right)^2\right\}$

圖 2 本文模型所採用的權重函數



它的具體解法在 2.2 節中展現。

同樣的，對區間模糊數的半徑，也可以進行類似的操作得到$\widehat{r_{x_0}}$。於是我們就得到了在x_0處對應變數的估計：$< \widehat{c_{x_0}}, \widehat{r_{x_0}} >$。接下來變換$x_0$的取值，使其取遍區間 $[\min\{x_1, \ldots, x_n\}, \max\{x_1, \ldots, x_n\}]$，重複以上過程，就可以得到對響應變數的連續估計。該步驟可通過程式實現並且並不困難,在 2.4 節中有討論相關性質。

2.2 回歸係數的估計

在本文模型中，回歸係數的估計用向量$\widehat{\beta_{x_0}}$表示，它的解法用如下定理展示：

定理 1　模糊數據的局部加權回歸模型中，回歸係數的估計和普通回歸模型類似，以處理中心點的回歸模式情況為例，對任意$x_0 \in [\min\{x_1,\dots,x_n\},\max\{x_1,\dots,x_n\}]$，可將回歸係數的估計寫成向量形式：$\widehat{\beta_{x_0}} = (X^T WX)^{-1}X^T \mathrm{W}\cdot \mathrm{c}$ 。其中

$$\mathrm{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathrm{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix},$$

$$\mathrm{W} = \begin{bmatrix} W_{x_0}(x_1) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & W_{x_0}(x_n) \end{bmatrix}$$

$$\mathrm{cost}(\hat{\beta}) = \sum_{i=1}^{n} W_{x_0}(x_i)\cdot(c_i - \widehat{\beta_{x_0}}_0 - \widehat{\beta_{x_0}}_1 \cdot x_i)^2$$

證明：一般回歸模型中的損耗函數（Cost Function）實際上是向量$\Delta \mathrm{c} = \mathrm{c} - \hat{c}$的長度（2-norm 長度）的平方，也就是$\| \Delta \mathrm{c} \|_2^2$。而局部加權模型里的損耗函數只是對向量Δc做了修正，如果令$\Delta c' = \sqrt{W}\cdot \Delta \mathrm{c}$,其中

$$\sqrt{W} = \begin{bmatrix} \sqrt{W_{x_0}(x_1)} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{W_{x_0}(x_n)} \end{bmatrix}$$

那麼局部加權模型里的損耗函數就能表示成和普通回歸模型里的損耗函數一樣的形式了，也就是說，$\mathrm{cost}(\hat{\beta}) = \| \Delta c' \|_2^2$。所以，使用和普通回歸模型一樣的方法求解回歸係數的估計。令$\frac{\partial}{\partial \hat{\beta}} cost(\hat{\beta}) = 0$, 即可得到：$\widehat{\beta_{x_0}} = (X^T WX)^{-1}X^T \mathrm{W}\cdot \mathrm{c}$

也就是說，$\widehat{c_{x_0}} = (1, x_0)\cdot(X^T WX)^{-1}X^T W\cdot c$

2.3 殘差分析

分析一個模型擬合效果的好壞的重要途徑之一是殘差分析，在普通回歸模型中，殘差大致可理解為應變數的實際值向量與擬合值向量的距離的平方。然而在模糊的框架下，如何定義這樣的距離，是存在爭議的。仿照 Diamond 對三角型模糊數之間距離的定義[6]，我們也可以同樣定義區間之間的距離。

定義 2 模糊回歸模式中的模糊殘差平方和 FSSE 為：

$$\mathrm{FSSE} = \mathrm{dist} < \mathrm{y}, \hat{y} >^2 = \sum_{i=1}^{n} [(\hat{c_i} - c_i + \hat{r_i} - r_i)^2 + (\hat{c_i} - c_i - \hat{r_i} + r_i)^2]$$

這個模糊殘差的定義的直觀理解就是著重模糊數的上下端點（或多個端點，比如三角型與梯形模糊數），分別平方求和。這樣我們就得到了一個實數來代表模糊回歸模型中的殘差，而實數可以用來比大小與作圖，從而可以使用經典情況下的殘差分析過程來做模糊回歸模型中的殘差分析。

2.4 數據模擬

本文所構建的模型並不僅限於自變數是一維的情況，但是為了更好的對該模型的效果有一個直觀的認識，在自變數是一維的情況下構建人工數據來進行模擬。這樣也方便用圖像來觀察。

給出樣本如下表 1：

表 1 數據模擬所用的人工數據

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
c	0.74	1.31	2.11	3.47	4.02	5.51	8.63	9.89	11.64	12.19
r	0.22	0.22	0.22	0.27	0.23	0.28	0.23	0.19	0.28	0.30
i	11	12	13	14	15	16	17	18	19	20
c	12.62	12.70	15.91	16.99	17.50	17.85	18.16	18.61	18.98	20.43
r	0.20	0.32	0.26	0.26	0.31	0.24	0.31	0.37	0.27	0.35
i	21	22	23	24	25	26	27	28	29	30
c	20.58	21.16	24.76	26.13	26.68	27.10	27.64	27.70	29.04	29.82
r	0.23	0.28	0.29	0.30	0.37	0.31	0.35	0.25	0.42	0.31

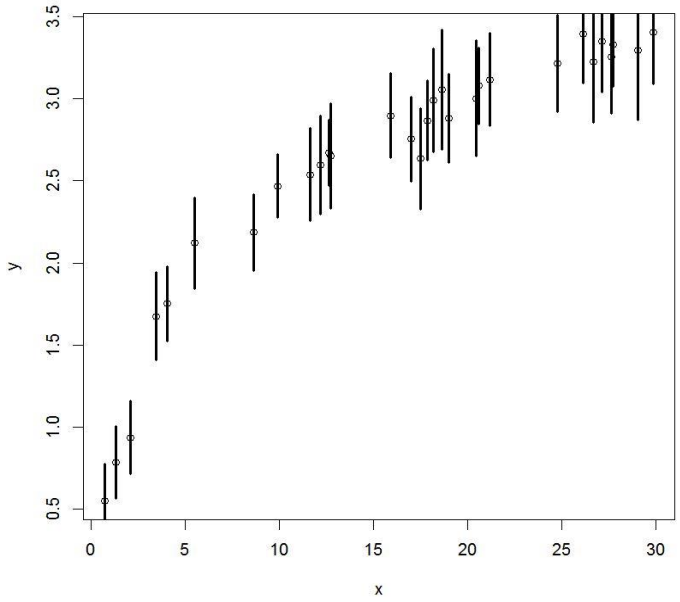
將這 30 個样本以圖像呈現如圖 3：

可以直觀地看出，該樣本自變數和因變數的大概關係並不是很理想的直線關係，而是更接近一種類似對數函數的關係。如果直接使用最簡單的線性回歸方式，按上下端點的值分別做線性回歸，得到兩條回歸直線：

$$y_u = 1.540317 + 0.083640 \cdot x$$

$$y_d = 1.102164 + 0.075895 \cdot x$$

圖 3 數據模擬中人工數據的圖示



計算可得，按照上下端點做簡單線性回歸所得到的結果的模糊殘差平方和 FSSE1 為 5.643559。而該方法擬合的直觀效果見圖 4。

可以看出，擬合效果差強人意。接下來使用本文構造的模型進行擬合，繪出擬合圖如圖 5。直觀來看，擬合效果比圖 4 的擬合效果好很多。同樣也可以通過計算得到新方法所得的模糊殘差平方和 FSSE2 為 0.6708636。

FSSE1 與 FSSE2 有數量級上的差距，也即新方法的擬合效果的確要好得多。

接下來對模型進行一個簡單的靈敏度分析。我們按一定比例變動某個模糊因變數的值，來觀察變動前后模型擬合效果的圖像有多大幅度的變化。

令第十個因變數的中心點變為原來的約三分之二，即從 2.5975319 變為 1.73，此時第十個樣本便顯得和其他樣本“格格不入”。如果它是一個異常值，那麼就應該從樣本中剔除出去，這也是很多情況下對原始樣本做的預處理。但是，為了體現“進入模型的每一個樣本帶來的信息都不能忽略”的觀念，以及在模擬數據的情況下觀察新模型的性質，把改變之後的樣本代入模型進行運算，并畫出擬合效果圖像如圖 6。

可以看出，在上述情況下模型對某個樣本的變動並不非常敏感，具有一定的穩定性。

圖 4 簡單按照區間上下端點做線性回歸的擬合效果示意圖

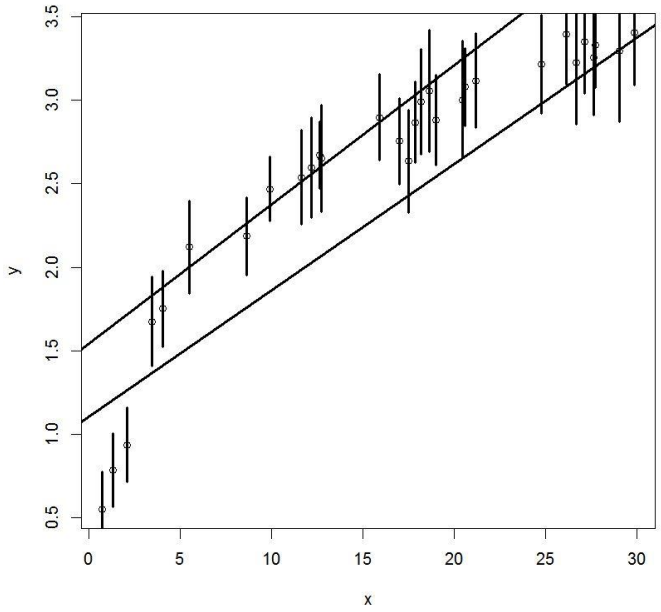
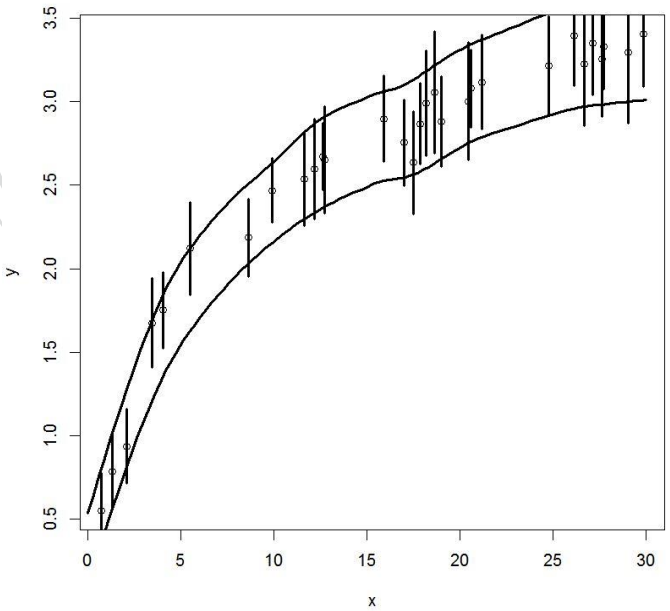
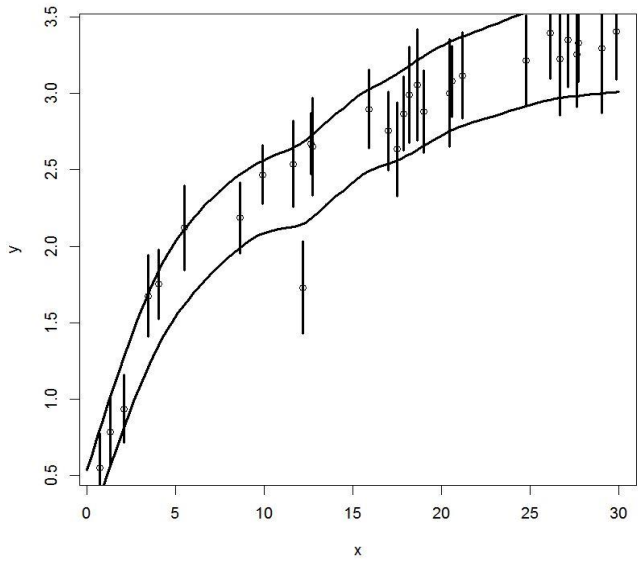


圖 5 模糊數據的局部加權方法擬合效果示意圖



模糊回歸分析和經典回歸分析的最大不同之一就是，經典回歸模型里假定不確定性來自一個獨立同分配（i.i.d.）的誤差，而模糊回歸模型則認為模型的不確定性來自多重隸屬度。很直觀的一個理解就是在模糊回歸模型里，應變數都是模糊數，這就是最好的體現。而自變數，在本文的模型里它是實數，實數類型的自變數加上存在模糊情況的應變數，是最典型的模糊回歸的模型構成。當然，也有

圖 6 添加一個異常值之後模糊數據的局部加權回歸方法擬合效果示意圖



自變數也是模糊數的情況，本文的構想仍然可以使用，只不過需要對模糊數與模糊數間的距離有一個很好的定義。另一個方面，從擬合效果上看，本文所提倡的局部加權方法要比普通最小平方法要好得多。但是，本文的方法也有其局限性。比如，該方法需要資料相對的“密集”，否則擬合效果一般。另外，局部加權方法容易造成過擬合的後果，直觀地說就是擬合曲線過於“彎曲”，以至於失去了宏觀分析自變數與應變數之間的關係。所以，在使用本文模型的時候，要注意權重函數不應選的過於“陡峭”。

3.實證分析

為了更好地了解新模型的實際效用，下面給出一個新模型的實證分析：一個利用模糊數對產品作評價的評價系統。如表 2 和表 3 所示，這是一個樣本容量為 50 的數據集[7]，其中自變數是 8 維的實數數據，代表不同型號汽車的 8 個參數，分別是汽車的價格、排氣量、馬力、峰值速度、加速度、市區內行駛油耗、市區外行駛油耗以及每公里損耗折現；因變數是專家團針對這 50 種車的這 8 個指標搭配給出的評價（最差，很差，差，中下，平均，中上，好，很好，最好），即專家認為該車型這幾個參數的搭配（包括價格），這樣的一個組合的合理程度。應該注意到的是，並沒有現成的體系來衡量一種車型在這 8 個方面的搭配是否合理，而專家也很難在沒有明確的量化指標的情況下給出一個確定的好壞程度的數值，所以借由語義變數（好，中等，差，等等）來表達專家們依據以往經驗給出的判斷。我們的目標是：第一，借由專家們的對這 50 個車型的判斷，在給出另一款新車型的各種參數搭配的時候，能夠不再次煩請專家團討論，量化衡量該車型搭配的合理與否；第二，分析出這 8 個參數對最後的評價的影響大小。這裡有兩點需要說明：第一，在“平均”與“中下”之間是有空間存在的，一個搭配很可能好於“中下”而不及“平均”，所以模型運行后給出的評價不能只是“不連續”的語義變數，而是需要量化結果，這樣才能達成“連續”的評定；第二，傳統的實數量化並不一定能很好的达到一些目的，比如若是有第 51、52 種車型，模型都給出“好”的評級，但是模型認為把 51 號評定為“中上”也不過分，而完全不能接受把 52 號放到“中上”的分類里去，如果想要得到這種並不是“非 0 即 1”的二元邏輯的結果，則很難由使用傳統實數量化的評級來實現。所以，我們採用文獻[8]中的方法，將語義變數轉化成為帶有隸屬度的模糊數（表 3 和圖 7）進行模型建構。

另外，表 3 中的模糊數遵循參考文獻中的表示方法。即所有的模糊數都是梯形模糊數，第 1,2 個數值代表梯形模糊數隸屬度為 1 時的區間的端點，第 3 個數值代表梯形模糊數第一端點與第二端點的距離，第 4 個數值代表梯形模糊數第三端點與第四端點的距離。圖 7 有直觀的解釋。

如果樣本數據有數量級上較大的差距，那麼有可能導致模型運行效果有較大

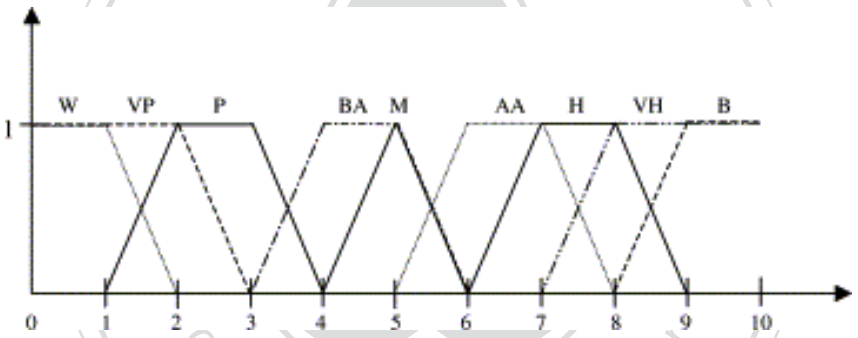
表 2 實證分析所用的數據集

	X1	X2	X3	X4	X5	X6	X7	X8	Y
	Price	Displacement	Potential	Speed	Acceleration	Fuel consumption		Cost/km	Experts
					0-100km/h	Urban	Extra		decision
	€	(cm3)	HP	(Km/h)	(s)	(Km/l)	(Km/l)	€	
1	21330	1598	120	200	10.5	8.8	15.6	0.41	AA
2	29864	1781	150	222	8.9	8.8	15.6	0.5	VH
3	26830	1895	118	206	10.4	8.8	16.7	0.49	AA
4	26004	1997	110	191	12.5	13.7	22.2	0.29	M
5	17613	1998	133	195	9	8.2	15.6	0.42	P
6	18120	1596	103	187	10.7	8.9	15.9	0.38	M
7	13170	1396	75	165	15.6	10.5	15.6	0.36	BA
8	19290	1997	136	200	9.6	7.9	14.9	0.44	P
9	26100	1998	150	210	9.6	7.2	13.3	0.48	M
10	29128	1988	155	215	9.5	7.5	12.8	0.52	VH
11	28715	1998	129	210	11	7.3	14.9	0.62	H
12	26494	1995	140	203	11.3	8.1	13.7	0.5	BA
13	22931	1997	136	208	10.8	8.7	15.4	0.51	AA
14	24248	1985	152	215	8.5	7.9	14.9	0.52	AA
15	19898	1595	105	192	11.3	8.6	15.5	0.41	AA
16	22200	1948	136	205	9.7	8.5	15.9	0.46	M
17	30320	1970	150	215	8.5	7.5	14.7	0.53	VP
18	19095	1390	75	173	12	16.7	12.2	0.43	AA
19	37390	2393	165	222	9.2	7.3	13.5	0.65	VH
20	63812	2771	193	232	10.1	5.7	11.9	0.88	H
21	32656	1781	180	228	7.4	9.2	15.9	0.55	M
22	54021	2793	193	228	8.6	7.3	12.7	0.84	H
23	20199	1997	90	175	14.5	14.3	21.7	0.3	H
24	15250	1596	103	180	11.5	9.6	16.9	0.37	BA
25	28379	1997	147	208	10	8.2	14.1	0.15	M
26	60942	3996	280	240	7.3	5.8	11.2	0.86	P
27	83666	3996	280	240	7.3	5.8	11.2	1.13	M
28	10750	1242	80	174	11.2	13.5	20	0.37	AA
29	66623	4293	281	250	6.7	5.7	11.2	1.01	VH
30	36772	1998	163	223	9.1	7.4	14.3	0.65	H
31	23235	1796	120	193	9	9.8	17.5	0.54	M
32	22176	1781	125	202	9.7	9.4	16.1	0.45	AA
33	40852	2171	170	226	9.1	8.2	14.1	0.63	M
34	37701	2446	129	201	12.1	9.2	16.9	0.39	P
35	22125	1998	133	206	10.2	7.7	14.1	0.46	P
36	12100	1242	60	155	14.3	13.7	20.8	0.3	AA
37	14530	1242	80	170	12.5	10.6	18.9	0.32	P
38	11078	1242	75	167	13.1	11.5	17.2	0.3	AA
39	15597	1596	101	185	11	11	18.5	0.37	H
40	72562	3996	281	250	6.7	5.8	11.6	1	AA
41	32030	1998	220	243	7.3	6.8	12.3	0.61	AA
42	32660	1796	118	202	5.9	10.4	17.5	0.52	AA
43	20193	1598	102	182	10.8	10.4	18.9	0.4	AA
44	40619	2597	170	219	9.5	6.1	11.8	0.71	VH
45	64764	3199	224	240	8.2	5.8	12.2	0.92	H
46	93117	4966	306	250	6.5	5.3	11.4	1.23	VH
47	11104	1360	75	170	13.2	11.2	18.9	0.3	VH
48	76132	3387	300	280	5.2	5.8	11.8	1.03	VH
49	11336	1149	60	160	15	12.7	19.2	0.45	VH
50	15423	1390	75	171	13.5	11.8	18.9	0.34	VH

表 3 實證分析中語義變數和對應模糊數的關係

W=Worst	(0,1,0,1)
VP = Very poor	(0,2,0,1)
P = Poor	(2,3,1,1)
BA = Below average	(4,5,1,1)
M = Average	(5,5,1,1)
AA = Above average	(6,7,1,1)
H = High	(7,8,1,1)
VH = Very high	(8,10,1,0)
B = Best	(9,10,1,0)

圖 7 實證分析中語義變數與對應模糊數的關係



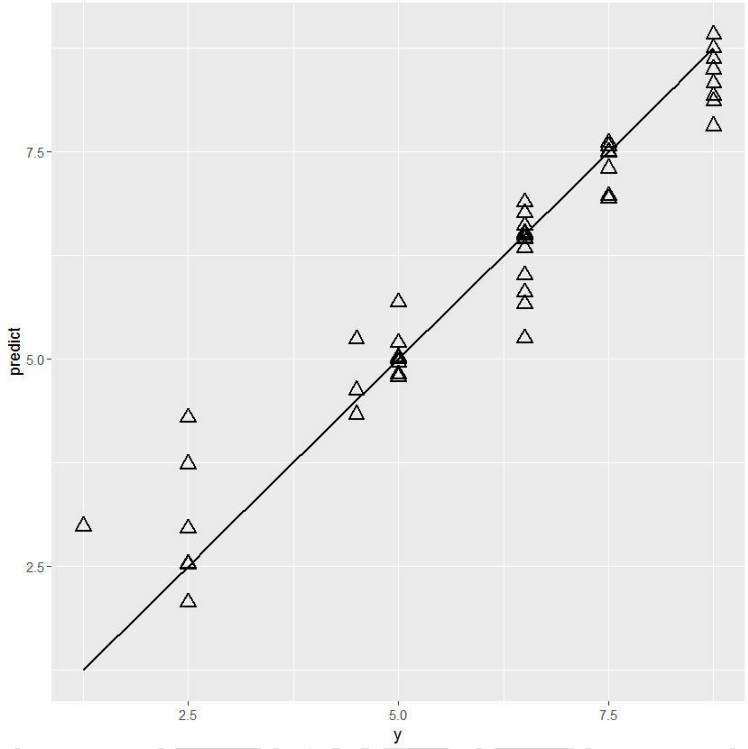
的偏差，所以先將樣本數據中心標準化，然後運用本文構建的模型進行擬合。首先對梯形模糊數中心點進行擬合，效果圖見圖 8。橫軸是 50 個模糊因變數中心點的實際數值，縱軸是運用本文方法對梯形模糊因變數中心點進行擬合的擬合值，直線是函數 y=x 的圖像。可以看出，大多數的點都落在直線的附近，表示中心點的擬合效果比較理想。

接下來對梯形模糊數的 4 個半徑做相同的處理得到相應的擬合值, 根據之前給出的定義，可以算得模糊殘差平方和 FSSE2 為 2.007249。同樣的，我們也可以直接對梯形模糊因變數的四個端點做簡單線性回歸, 根據定義我們也能得到一個模糊殘差平方和 FSSE1，算得其值為 582.5381。所以，從模糊殘差平方和的角度來看，新模型的擬合效果的確不錯。

可以看出，樣本中因變數是梯形模糊數，模型擬合后得到的也是梯形模糊數。

但是具體到我們的第一個目標，給出某車型的 8 個參數，我們得到的擬合結果是

圖 8 模糊數據的局部加權回歸對模糊因變數中心點的擬合效果



一個梯形模糊數，并不能直接用它來對該車型做出評價。我們需要將得到的擬合結果反模糊化。假設我們得到的擬合結果是$<\hat{c},\hat{r_1},\hat{r_2},\hat{r_3},\hat{r_4}>$，容易得知，此時已經不一定有$\hat{r_1}+\hat{r_2}+\hat{r_3}+\hat{r_4}=0$的結論，所以擬合結果的反模糊化值（也就是擬合出來的模糊數的中心點）為：

$$\hat{c'}=\hat{c}+\hat{r_1}+\hat{r_2}+\hat{r_3}+\hat{r_4}$$

此時，我們選擇把$\hat{c'}$作為最終的量化評判結果，通過對照表 3 或圖 7 進行判斷。比如，若是擬合結果的中心點為 2.7，我們認為該車型的搭配屬於“差”的隸屬程度為 1，屬於“很差”的隸屬程度為 0.3；若是擬合結果的中心點為 7.5，那麼認為該搭配屬於“好”的隸屬程度為 1，屬於“中上”和“很好”的隸屬程度都為 0.5。

此外，值得注意的是，上文所做的實際上是解決了一個分類問題，即把一個關於車型的參數搭配的輸入歸類到“平均”、“中上”等八類中去。並且，和普通的分類方法（比如邏輯回歸方法）不同的是，我們得到的輸出是該輸入屬於各個類別的隸屬度，而並非輸出某個具體的類或該輸入屬於某個類的幾率。具體地說，某輸入屬於類 A 和類 B 的隸屬度為 0.1、0.15 的情況和隸屬度為 0.8、0.85

的情況完全體現了兩種不同信息—儘管最後的選擇都是歸為 B 類；而這樣的信息是普通的分類法，即去得到某個輸入屬於類 A 和類 B 的幾率是多少的方法所不能體現的，因為有屬於 A、B 的幾率之和為 1 的約束。這也詮釋了模糊理论對破除“非零即一”的二元邏輯的貢獻。

接下來我們來實現第二個目標：分析這 8 個因素對最終結果的影響大小。若是簡單的線性回歸，我們只要看各個係數的大小就能知道各個因素的影響大小，但是本模型採用的是基於線性回歸的局部加權回歸方法。從其理論構造容易得知，本文模型的各個係數並不是一個常數，而是一個對自變數位置的函數。我們的樣本數量有 50 個，所以在每個樣本處，對中心點和 4 個半徑，都有相應的 9 個係數（第一個係數是常數項），其數據量比較大，示意圖列表如表 4。

表 4 實證分析中各個回歸係數的表示方法示意圖

i	1					…	50				
	c	r1	r2	r3	r4	…	c	r1	r2	r3	r4
常數項	A11c	A11r1	A11r2	A11r3	A11r4	…	A501c	A501r1	A501r2	A501r3	A501r4
…			…			…			…		
Cost/km	A19c	A19r1	A19r2	A19r3	A19r4	…	A509c	A509r1	A509r2	A509r3	A509r4

其中$\mathrm{Aijc;i}=1,\ldots,50;\mathrm{j}=1,\ldots 9$表示對中心點做回歸時對第 j 個自變數在第 i 個樣本時的係數的值。$\mathrm{Aijr1,Aijr2,Aijr3,Aijr4}$的含義以此類推。就以第 9 個自變數 Cost/km 來說，它對中心點的回歸的係數有 50 個，為了防止極端值對平均值造成過大影響，採用這 50 個係數的中位數$\mathrm{Am9c}$來代表 Cost/km 這第 8 個自變數對擬合結果中心點的影響大小。即：

$$\mathrm{Am9c}=\mathrm{median}\{\mathrm{A19c},\ldots,\mathrm{A509c}\}$$

同樣可以得到 Cost/km 因素對四個半徑的擬合“影響大小”$\mathrm{Am9r1,Am9r2,Am9r3,Am9r4}$。但是，上文已經提到，作者認為中心點作為“位置參數”，其重要性要高於半徑，所以最終 Cost/km（算上常數項后的第 9 個因變數）對評價結果的影響因子定義為：

$$IF_9=|\mathrm{Am9c}|+\log(1+\mathrm{mean}(|\mathrm{Am9r1}|,|\mathrm{Am9r2}|,|\mathrm{Am9r3}|,|\mathrm{Am9r4}|))$$

其他幾個因變數的影響因子也以此類推。

接下來我們把數據代入，得到最後的數值結果列表如表 5。

表 5 各個因變數對最終評判的影響因子

j	1	2	3	4	5	6	7	8	9
Amjc	3.27	-3.16	0.54	-8.64	7.62	0.56	2.83	-0.71	2.41
Amjr1	-1.22	0.49	-0.30	1.00	-0.64	0.04	-0.19	0.13	-0.25
Amjr2	-0.26	0.20	-0.18	1.01	-0.55	0.13	0.07	0.02	-0.15
Amjr3	0.17	-0.51	0.57	-1.71	1.60	0.01	0.08	-0.07	0.41
Amjr4	1.33	-0.13	0.01	-0.59	0.13	-0.07	0.06	-0.06	0.10
IFj	3.82	2.88	0.78	7.91	8.17	0.62	2.91	0.64	2.62

由此可見，除去常數項之後，8 個因素裡面對最終評定的影響最大的两个是：峰值速度和馬力大小；市區耗油量，價格以及每公里損耗折現值也有較大的影響力；加速度，排氣量和市區外耗油對該車型搭配合理與否的最終評定的影響不大。

4.結語

模糊回歸分析一直是模糊統計領域的熱門研究方向。由實證分析可以看出，本文構造的新模型運用在實際問題中是比較靈活的，尤其是把中心點和半徑分開討論的思路，可以在實際使用中根據需求做出調整；并且，它首次將局部加權的思想融入到模糊回歸分析領域，它有很顯著的優點，比如擬合效果好，在數據量大、密集的情況下效果尤佳，這與當今“大數據”的時代背景剛好吻合；也有它非常獨特的地方，比如回歸係數的估計不是一個確定值，而是一個關於關注點位置的函數，關注點有所移動，得到的係數的估計就有所不同；另外，也有一些不足之處，那就是模型需要的計算量略顯偏大，不過這在當今電腦計算能力大大提升的背景下，並不是一個難以解決的問題。總之，本文從一個新的著手點，為模糊回歸領域的研究提供了一條新的思路。

參考文獻:

[1] L.A. Zadeh, Fuzzy sets, Information and Control, Volume 8, Issue 3, June 1965, pp.338–353

[2] H. Tanaka, S. Uejima, K. Asai, Linear regression analysis with fuzzy model, IEEE Trans. Sys., Man. Cyber., 12 (1982), pp. 903–907.

[5] William S. Cleveland, Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association, Vol. 74,No. 368.(Dec., 1979),pp. 829-836.

[6]Phil Diamond, Fuzzy Least Squares, Information Sciences 46(3), 1988, pp.141 -157

[7] Pierpaolo D'Urso, Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data ,Computational Statistics & Data Analysis, Volume 42, Issues 1–2, (2003), pp.47–72.

[8] P. Anand Raj, D. Nagesh Kumar, Ranking alternatives with fuzzy weights using maximizing set and minimizing set ,Fuzzy Sets and Systems,1999,pp365-375

[3]吳柏林，模糊統計導論第二版（2015），五南出版社（台北），p153.

[4]陳孝煒、吳柏林，區間回歸與模糊樣本分析，管理科學與統計決策，4(1)，2007