رگرسیون فازی مبتنی بر کمترین قدرمطلق انحرافات

مریم کلکین نما[۱]، سید محمود طاهری[۱]

m − kelkinnama@math.iut.ac.ir, taheri@cc.iut.ac.ir

چکیده

در این مقاله به مدل رگرسیونی با متغیرهای ورودی دقیق و خروجی فازی پرداخته شده است. رویکرد پیشنهادی به این مساله، استفاده از روش کمترین قدرمطلق خطا است که با معرفی و استفاده از یک متر در محیط اعداد فازی صورت گرفته است. به‌علاوه، یک شاخص تطبیق برای بررسی نیکویی برازش مدل‌های رگرسیونی فازی معرفی شده است. روش پیشنهادی با مثال عددی تشریح شده و نشان داده شده است که در مدل‌سازی داده‌های شامل داده پرت عملکرد خوبی دارد.

واژه‌های کلیدی: رگرسیون فازی، کمترین قدرمطلق انحرافات، نیکویی برازش، داده پرت، فضای متری.

۱ مقدمه

از زمان ارائه‌ی نظریه مجموعه‌های فازی توسط پروفسور زاده، استفاده از این نظریه در بخش‌های وسیعی از علم آمار در حال گسترش است. یکی از مهم‌ترین و کاربردی‌ترین این بخش‌ها، رگرسیون است.

اولین بار تاناکا و همکاران[۱۵] در سال ۱۹۸۲ موضوع رگرسیون فازی را مطرح کردند. آن‌ها مدل رگرسیون خطی با مشاهدات دقیق و پارامترهای فازی را مورد توجه قرار دادند. در رهیافت آن‌ها، که به‌نام «رگرسیون امکانی» نیز شناخته شده است، برازش مدل رگرسیون خطی فازی به‌صورت مینیمم کردن مجموع ابهام در مقدار برآورد شده مشاهدات انجام می‌شود با توجه به این قید که میزان عضویت مقدار مشاهده شده خروجی (دقیق) در مقدار برآورد شده‌ی فازی متناظر آن، حداقل به میزان h باشد. این مساله عموماً معادل با یک مساله برنامه‌ریزی خطی (و گاهی غیرخطی) می‌شود که با حل آن، مدل بهینه رگرسیون فازی به‌دست می‌آید.

رهیافت دیگر در زمینه رگرسیون فازی توسط دایاموند[۹] و کلمینس [۶] در سال ۱۹۸۷ ارائه شد که با استفاده از روش کمترین مربعات به بررسی و برازش مدل رگرسیون فازی می‌پردازد و می‌توان آن‌را تعمیم یافته رگرسیون کمترین مربعات معمولی دانست. مبنای این روش، استفاده از یک فاصله روی مجموعه اعداد فازی است که براساس آن، مجموع مربعات فواصل مقادیر خروجی فازی مشاهده شده و مقادیر برآورد فازی آن‌ها، مینیمم می‌شود.

تاکنون افراد مختلفی به بررسی انواع روش‌های رگرسیون امکانی و رگرسیون کمترین مربعات، تحت شرایط و حالت‌های مختلف، پرداخته‌اند. چند نمونه از مطالعات جدید در مورد رگرسیون امکانی را می‌توانید در [۱۲] و [۱۴] ببینید. هم‌چنین برای مطالعه چند تحقیق اخیر درباره رگرسیون کمترین مربعات فازی به [۵] و [۱۱] مراجعه کنید. منابع بیشتر را همراه با مرور برخی تحقیقات می‌توانید در [۳] بیابید.

یکی از موضوعات مهم در رگرسیون، مدل‌سازی مجموعه

[۱] دانشکده علوم ریاضی، دانشگاه صنعتی اصفهان

داده‌هایی است که در آن‌ها یک یا چند داده پرت وجود دارد. رویکرد متداول در رگرسیون آماری، استفاده از روش کمترین مربعات است که به‌ویژه زمانی که توزیع خطاها نرمال باشد، دارای خواص بهینه مناسبی است. اما زمانی که در مجموعه داده‌ها، داده(ها)ی پرت که ناشی از توزیع‌های خطا با دم سنگین هستند، وجود داشته باشند، معمولاً اثرات نامناسبی بر برآوردهای حاصل از روش کمترین مربعات خواهند گذاشت. در این موارد می‌توان از روش‌های نیرومند (استوار) استفاده کرد تا اثر داده‌های پرت را بر برآوردهای ضرایب کمتر کند. یکی از روش‌های نیرومند، روش کمترین قدرمطلق خطا ($Least\ Absolute\ Deviations:\ LAD$) است. این روش به‌خصوص زمانی که توزیع خطا دم سنگین داشته باشد، کارایی بالایی داشته و به‌طور وسیعی مورد استفاده قرار گرفته است [۱۰].

در عمل ممکن است در مجموعه‌ی داده‌های فازی نیز داده‌ی پرت وجود داشته باشد. لذا بررسی روش‌های رگرسیون فازی نیرومند نیز ضروری می‌نماید. به‌عنوان تعمیمی از رگرسیون کمترین قدرمطلق خطای معمولی، برخی محققان به بررسی رگرسیون کمترین قدرمطلق خطای فازی ($FLAD$) نیز پرداخته‌اند. چانگ و لی [۷] در سال ۱۹۹۴ با استفاده از یک روش رتبه‌بندی اعداد فازی، به مینیمم کردن قدرمطلق تفاضل رتبه اعداد فازی مشاهده شده و اعداد فازی برآورد شده متناظر آن‌ها پرداختند. ترابی و بهبودیان [۱۶] در سال ۲۰۰۷ روشی را برای انجام رگرسیون کمترین قدرمطلق انحرافات فازی با استفاده از اصل تجزیه برای مدلی با ورودی و خروجی فازی ارایه نمودند. چوی و باکلی [۸] در سال ۲۰۰۸ به بررسی دو مدل رگرسیون فازی به دو شیوه مختلف پرداختند. در مدل با متغیر ورودی غیرفازی و متغیر خروجی فازی، قدرمطلق تفاضل کران‌های پایین (و بالای) $\alpha-$ برش‌های اعداد فازی مشاهده شده و برآورد آن‌ها را مینیمم کردند و برای مدل با ورودی و خروجی فازی ابتدا با غیر فازی‌سازی همه متغیرها، رگرسیون کمترین قدرمطلق خطای معمولی را اجرا کرده و سپس با تعریف یک عبارت خطای فازی، قدرمطلق فاصله بین اعداد فازی حاصل را مینیمم کردند.

آن‌چه که در این مقاله صورت گرفته است، ارایه یک متر جدید مبتنی بر قدرمطلق در مجموعه اعداد فازی است و بر اساس آن روشی برای انجام رگرسیون کمترین قدر مطلق انحرافات برای مدل با ورودی دقیق و خروجی فازی ارایه شده است. سپس با معرفی یک شاخص تطبیق و با به‌کارگیری آن در یک مثال عددی به بررسی این روش رگرسیونی پرداخته شده است. نتیجه شده است که روش پیشنهادی برای مدل‌سازی مجموعه داده‌های شامل داده(های) پرت، عملکرد مناسبی دارد.

۲ مفاهیم اولیه

فرض کنید X یک مجموعه مرجع باشد. یک مجموعه فازی $\tilde{A}$ از X توسط تابع عضویت آن $\tilde{A}: X \longrightarrow [0,1]$ تعریف می‌شود. مجموعه $\alpha-$ برش $\tilde{A}$ به‌صورت $0 < \alpha \leq 1$، $\tilde{A}_\alpha = \{x : \tilde{A}(x) \geq \alpha\}$ تعریف می‌شود.

تعریف ۱ هر مجموعه فازی نرمال از $\mathbb{R}$ را که $\alpha-$برش‌های آن به‌صورت بازه‌های بسته باشد، عدد فازی می‌نامیم.

تعریف ۲ عدد فازی $\tilde{M}$ را عدد فازی LR گوییم اگر اعداد حقیقی m ، $\alpha > 0$، $\beta > 0$ موجود باشند به‌طوری که

$$M(x) = \begin{cases} L\left(\frac{m-x}{\alpha}\right), & x < m \\ R\left(\frac{x-m}{\beta}\right), & x \geq m \end{cases}$$

که برای $x \leq 0$، $L(x) = R(x) = 1$ و برای $x \geq 1$، $L(x) = R(x) = 0$.

در این صورت $\tilde{M}$ را به‌صورت $\tilde{M} = (m, \alpha, \beta)_{LR}$ نمایش می‌دهیم که m مرکز، α و β به‌ترتیب پهنای چپ و راست نامیده می‌شوند. اگر $L(x) = R(x) = 1 - x$، آن‌گاه $\tilde{M}$ را عدد فازی مثلثی می‌نامیم و آن را با $\tilde{M} = (m, \alpha, \beta)_T$ نشان می‌دهیم به‌ویژه، اگر $\alpha = \beta$، آن گاه $\tilde{M}$ را عدد فازی مثلثی متقارن نامیده و به‌صورت $(m, \alpha)_T$ نشان می‌دهیم.

حساب اعداد فازی از جمله اعمال جبری مانند ضرب اسکالر و جمع و ضرب بر اعداد فازی، براساس اصل گسترش تعریف می‌شود. برای مطالعه بیشتر در این زمینه به [۱۸] مراجعه کنید. برای اهداف این مقاله قضیه زیر را در مورد ضرب اسکالر و جمع دو عدد فازی LR بیان می‌کنیم.

قضیه ۱ فرض کنید $\tilde{N}, \tilde{M}$ دو عدد فازی LR به‌صورت $\tilde{M} = (m, \alpha, \beta)_{LR}$ و $\tilde{N} = (n, \gamma, \delta)_{LR}$ و $\lambda \in \mathbb{R}$ باشند آن‌گاه داریم

$$\begin{aligned} \lambda \otimes \tilde{M} &= \tilde{M} \otimes \lambda \\ &= \begin{cases} (\lambda m, \lambda\alpha, \lambda\beta)_{LR}, & \lambda > 0 \\ (\lambda m, -\lambda\beta, -\lambda\alpha)_{RL}, & \lambda < 0 \end{cases} \end{aligned}$$

$$\begin{aligned} \tilde{M} \oplus \tilde{N} &= (m, \alpha, \beta)_{LR} \oplus (n, \gamma, \delta)_{LR} \\ &= (m + n, \alpha + \gamma, \beta + \delta)_{LR} \end{aligned}$$

۳ رگرسیون کمترین قدرمطلق انحرافات فازی

مدل رگرسیونی $\mathbf{y} = \mathbf{X}\beta + \epsilon$ را درنظر بگیرید که در آن $\mathbf{y}$ بردار مقادیر متغیر وابسته، $\mathbf{X}$ ماتریس مشاهدات متغیرهای مستقل، β بردار ضرایب مجهول و ϵ برداری از ϵ_iها، یعنی خطای بین مقدار مشاهده شده y_i و مقدار برآورد آن است. روش متداول برای برآورد ضرایب این مدل رگرسیونی، استفاده از روش کمترین مجموع مربعات خطا است. این روش به‌علت راحتی محاسبات و داشتن ویژگی‌های مطلوب به‌ویژه در مواقعی که توزیع خطاها نرمال باشد، بسیار مورد استفاده قرار می‌گیرد. اما هنگامی که در مجموعه داده‌ها، یک یا چند داده پرت وجود داشته باشد، ممکن است استفاده از روش کمترین مربعات روش مناسبی نباشد، چرا که نتایج تحلیل‌ها ممکن است به‌شدت تحت تاثیر داده‌های پرت قرار گیرد. یک راه برای کاهش اثر این گونه داده‌ها، استفاده از معیار قدرمطلق خطا به‌جای مجذور خطا است .

برآوردگرهای روش‌های کمترین قدرمطلق و کمترین مربعات خطاها برای برآورد ضرایب مدل رگرسیونی، عضو رده برآوردگرهای L_p هستند که توسط مینیمم کردن آن‌چه که به‌نام متر مینکوسکی [۲] یا L_P- نرم معروف است

$$\{\sum |\epsilon_i|^p\}^{1/p} \quad p \geq 1$$

به‌دست می‌آیند. اگر $p = 1$، متر قدرمطلق (L_1- نرم) را خواهیم داشت و مینیمم‌سازی این فاصله، روش رگرسیونی کمترین قدرمطلق خطا نامیده می‌شود. اگر $p = 2$، این معیار به متر اقلیدسی تبدیل می‌شود که مینیمم‌سازی مبتنی بر آن، منجر به روش کمترین مربعات خطا می‌شود [۱۰].

در عمل ممکن است در مجموعه داده‌های فازی نیز، داده پرت وجود داشته باشد. لذا شاید مناسب‌تر باشد که برای مدل‌سازی فازی داده‌های شامل داده پرت نیز از رگرسیون قدرمطلق انحرافات استفاده کنیم.

در این بخش به معرفی یک فاصله بین اعداد فازی می‌پردازیم که مبتنی بر قدرمطلق بیان می‌شود و پس از بررسی ویژگی‌های آن، رگرسیون کمترین قدرمطلق انحرافات فازی را بر اساس این فاصله انجام می‌دهیم.

[۲] Minkowsky

۱.۳ معرفی یک متر جدید برای اعداد فازی

تعریف ۳ فرض کنید $\tilde{X}=(m_x,\alpha_x,\beta_x)_{LR}$ و $\tilde{Y}=(m_y,\alpha_y,\beta_y)_{LR}$، دو عدد فازی LR باشند، آن‌گاه فاصله بین دو عدد فازی $\tilde{X}$ و $\tilde{Y}$ به‌صورت زیر تعریف می‌شود

$$D_{LR}(\tilde{X},\tilde{Y})=|m_x-m_y|+|(m_x-l\alpha_x) -(m_y-l\alpha_y)|+|(m_x+r\beta_x)-(m_y+r\beta_y)| \qquad (۱)$$

که در آن $r=\int_0^1 R^{-1}(w)dw$ و $l=\int_0^1 L^{-1}(w)dw$.

این فاصله بین اعداد فازی در تمام خواص یک متر صدق می‌کند.

قضیه ۲ اگر $\mathcal{F}_{LR}(\mathbb{R})$ مجموعه همه اعداد فازی LR از مجموعه اعداد حقیقی $\mathbb{R}$ باشد، آن‌گاه $(\mathcal{F}_{LR}(\mathbb{R}),D_{LR})$ یک فضای متری است.

اثبات کافی است نشان دهیم که چهار ویژگی زیر برقرار است

۱) $\forall\tilde{X},\tilde{Y}\in\mathcal{F}_{LR},\ \ D_{LR}(\tilde{X},\tilde{Y})\geq 0$

۲) $D_{LR}(\tilde{X},\tilde{Y})=0\ \Leftrightarrow \tilde{X}=\tilde{Y}$

۳) $\forall\tilde{X},\tilde{Y}\in\mathcal{F}_{LR},\ \ D_{LR}(\tilde{X},\tilde{Y})=D_{LR}(\tilde{Y},\tilde{X})$

۴) $\forall\tilde{X},\ \tilde{Y},\tilde{Z}\in\mathcal{F}_{LR},$

$$D_{LR}(\tilde{X},\tilde{Z})\leq D_{LR}(\tilde{X},\tilde{Y})+D_{LR}(\tilde{Y},\tilde{Z})$$

شرط‌های اول و دوم به‌وضوح برقرار هستند، برای اثبات شرط سوم توجه کنید که

$$D_{LR}(\tilde{X},\tilde{Y})=|m_x-m_y|+|(m_x-l\ \alpha_x)-(m_y-l\ \alpha_y)|+|(m_x+r\ \beta_x)-(m_y+r\ \beta_y)|$$
$$=|m_y-m_x|+|(m_y-l\ \alpha_y)-(m_x-l\ \alpha_x)|+|(m_y+r\ \beta_y)-(m_x+r\ \beta_x)|=D(\tilde{Y},\tilde{X})$$

برای بررسی شرط چهارم نیز با استفاده از نامساوی مثلثی داریم

$$|m_x-m_z|\leq|m_x-m_y|+|m_y-m_z|$$
$$|(m_x-l\ \alpha_x)-(m_z-l\ \alpha_z)|\leq|(m_x-l\ \alpha_x)-(m_y-l\ \alpha_y)|+|(m_y-l\ \alpha_y)-(m_z-l\ \alpha_z)|$$
$$|(m_x+r\ \beta_x)-(m_z+r\ \beta_z)|\leq|(m_x+r\ \beta_x)-(m_y+r\ \beta_y)|+|(m_y+r\ \beta_y)-(m_z+r\ \beta_z)|$$

با جمع کردن طرف‌های چپ این سه نامساوی و نیز طرف‌های راست آن‌ها، ملاحظه می‌شود که شرط چهارم نیز برقرار است.□

قضیه ۳ $(\mathcal{F}_{LR}(\mathbb{R})$ و $D_{LR})$ یک فضای متری کامل است.

اثبات فرض کنید $\{\tilde{X}_n=(m_{x_n},\alpha_{x_n},\beta_{x_n})_{LR};n\geq 1\}$ یک دنباله کوشی در $\mathcal{F}_{LR}(\mathbb{R})$ باشد. اگر $i,j\to\infty$، داریم $|m_{x_i}-m_{x_j}|\leq D_{LR}(\tilde{X}_i,\tilde{X}_j)\to 0$. لذا دنباله $\{m_{x_n}\}_{n=1}^{\infty}$ یک دنباله کوشی در $\mathbb{R}$ است. در نتیجه با توجه به کامل بودن $\mathbb{R}$، $m_{x_n}\to m_x$.

نیز اگر $i,j\to\infty$ آن‌گاه

$$|(m_{x_i}-l\alpha_{x_i})-(m_{x_j}-l\alpha_{x_j})|\leq D_{LR}(\tilde{X}_i,\tilde{X}_j)\to 0$$

پس دنباله $\{m_{x_n}-l\alpha_{x_n}\}_{n=1}^{\infty}$ یک دنباله کوشی در $\mathbb{R}$ می‌باشد. چون $\{m_{x_n}\}_{n=1}^{\infty}$ یک دنباله کوشی و l یک عدد ثابت است، نتیجه می‌شود که $\{\alpha_{x_n}\}_{n=1}^{\infty}$ نیز یک دنباله کوشی است. پس $\alpha_{x_n}\overset{n\to\infty}{\longrightarrow}\alpha_x$. به‌طور مشابه می‌توان ثابت کرد که $\beta_{x_n}\overset{n\to\infty}{\longrightarrow}\beta_x$.

تمام موارد بالا نتیجه می‌دهند که $D_{LR}(\tilde{X}_n,\tilde{X})\to 0$. پس $\tilde{X}_n\overset{n\to\infty}{\longrightarrow}\tilde{X}$ که $\tilde{X}=(m_x,\alpha_x,\beta_x)_{LR}$.□

۲.۳ رگرسیون فازی مبتنی بر کمترین قدرمطلق انحرافات

فرض کنید مجموعه داده‌های $\{(\mathbf{x}_i,\tilde{y}_i),\ i=1,\cdots,n\}$ را در اختیار داریم که در آن $\mathbf{x}_i=(x_{i0},x_{i1},\cdots,x_{ip})$ با $x_{i0}=1$، بردار مشاهدات ورودی غیرفازی (دقیق) و غیرمنفی

$(x_{ij} \geq 0, \forall i, j)$ و $\tilde{y}_i = (m_{y_i}, \alpha_{y_i}, \beta_{y_i})_{LR}$، $i = 1, \cdots, n$، مقدار خروجی مشاهده شده فازی باشد.

اکنون می‌خواهیم مدل زیر را به این داده‌ها برازش دهیم

$$\begin{aligned} \tilde{Y}_i &= (\tilde{A}_0 \otimes x_{i0}) \oplus (\tilde{A}_1 \otimes x_{i1}) \oplus (\tilde{A}_2 \otimes x_{i2}) \\ & \oplus \cdots \oplus (\tilde{A}_p \otimes x_{ip}), \quad i = 1, \cdots, n \end{aligned}$$

که در آن $\tilde{A}_j = (m_j, \alpha_j, \beta_j)_{LR},\ j = 0, \cdots, p$ها، ضرایب مدل، به‌صورت اعداد فازی LR هستند.

در بین انواع اعداد فازی، اعداد فازی مثلثی به‌دلیل راحتی محاسبات، بیشتر استفاده می‌شوند. لذا ما نیز در ادامه $\tilde{A}_j, \tilde{y}_i$ها را اعداد فازی مثلثی در نظر می‌گیریم.

با استفاده از روابط موجود بین ضرب و جمع اعداد فازی LR در قضیه ۱، معلوم می‌شود که $\tilde{Y}_i$ یک عدد فازی مثلثی به‌صورت زیر خواهد بود

$$\tilde{Y}_i = (\sum_{j=0}^{p} m_j x_{ij}, \sum_{j=0}^{p} \alpha_j x_{ij}, \sum_{j=0}^{p} \beta_j x_{ij})_T, i = 1, \cdots, n.$$

برای برازش مدل رگرسیونی کمترین قدرمطلق خطای فازی، مجموع فاصله بین $\tilde{Y}_i$ و $\tilde{y}_i$ را بر اساس متر (۱) مینیمم می‌کنیم. یعنی

$$\begin{aligned} \sum_{i=1}^{n} D_T(\tilde{y}_i, \tilde{Y}_i) = \sum_{i=1}^{n} \{ &|m_{y_i} - \sum_{j=0}^{p} m_j\ x_{ij}| \\ &+ |(m_{y_i} - \frac{1}{2}\ \alpha_{y_i}) - (\sum_{j=0}^{p} m_j\ x_{ij} - \frac{1}{2}\ \sum_{j=0}^{p} \alpha_j\ x_{ij})| \\ &+ |(m_{y_i} + \frac{1}{2}\ \beta_{y_i}) - (\sum_{j=0}^{p} m_j\ x_{ij} + \frac{1}{2}\ \sum_{j=0}^{p} \alpha_j\ x_{ij})| \} \end{aligned}$$

برای این‌که بتوان مساله مینیمم‌سازی بالا را به‌صورت یک مساله برنامه‌ریزی خطی نوشت تا به‌راحتی قابل حل باشد، به شیوه زیر عمل می‌کنیم

با استفاده از متغیرهای غیرمنفی d_i^{m1} و d_i^{m2}، تعریف می‌کنیم $|m_{y_i} - \sum_{j=0}^{p} m_j\ x_{ij}| = d_i^{m1} + d_i^{m2}$ که $m_{y_i} - \sum_{j=0}^{p} m_j\ x_{ij} = d_i^{m1} - d_i^{m2}$.

به‌طور مشابه با استفاده از متغیرهای غیرمنفی d_i^{L1} و d_i^{L2} و d_i^{R1} و d_i^{R2} دو قدرمطلق دیگر را نیز به‌صورت زیر می‌نویسیم

$$|m_{y_i} - \frac{1}{2}\alpha_{y_i} - (\sum_{j=0}^{p} m_j x_{ij} - \frac{1}{2}\sum_{j=0}^{p} \alpha_j x_{ij})| = d_i^{L1} + d_i^{L2}$$

$$m_{y_i} - \frac{1}{2}\alpha_{y_i} - (\sum_{j=0}^{p} m_j x_{ij} - \frac{1}{2}\sum_{j=0}^{p} \alpha_j x_{ij}) = d_i^{L1} - d_i^{L2}$$

$$|m_{y_i} + \frac{1}{2}\beta_{y_i} - (\sum_{j=0}^{p} m_j x_{ij} + \frac{1}{2}\sum_{j=0}^{p} \beta_j x_{ij})| = d_i^{R1} + d_i^{R2}$$

$$m_{y_i} + \frac{1}{2}\beta_{y_i} - (\sum_{j=0}^{p} m_j x_{ij} + \frac{1}{2}\sum_{j=0}^{p} \beta_j x_{ij}) = d_i^{R1} - d_i^{R2}$$

سرانجام مساله مینیمم‌سازی بالا به‌صورت زیر صورت‌بندی می‌شود

$$Z = \sum_{i=1}^{n} (d_i^{m1} + d_i^{m2} + d_i^{R1} + d_i^{R2} + d_i^{L1} + d_i^{L2})$$

$s.t.$

$$m_{y_i} - \sum_{j=0}^{p} m_j\ x_{ij} = d_i^{m1} - d_i^{m2},$$

$$m_{y_i} - \frac{1}{2}\alpha_{y_i} - (\sum_{j=0}^{p} m_j x_{ij} - \frac{1}{2}\sum_{j=0}^{p} \alpha_j x_{ij}) = d_i^{L1} - d_i^{L2},$$

$$m_{y_i} + \frac{1}{2}\beta_{y_i} - (\sum_{j=0}^{p} m_j x_{ij} + \frac{1}{2}\sum_{j=0}^{p} \beta_j x_{ij}) = d_i^{R1} - d_i^{R2},$$

$$d_i^{m1},\ d_i^{m2},\ d_i^{R1},\ d_i^{R2},\ d_i^{L1},\ d_i^{L2} \geq 0,$$

$$\alpha_j \geq 0, \beta_j \geq 0,\ j = 0, \cdots, p.$$

مساله برنامه‌ریزی خطی بالا با استفاده از نرم‌افزارهایی مانند $GAMS$ و $LINGO$ قابل حل می‌باشد. با حل مساله فوق، مراکز و پهناهای اعداد فازی مثلثی $\tilde{A}_j$ ها (یعنی ضرایب فازی مدل) به‌دست می‌آیند.

۴ نیکویی برازش

در بررسی مدل‌های رگرسیون فازی، باید معیاری در دست داشت تا به‌وسیله آن، مدل‌های مختلف را با هم مقایسه کرد. در واقع باید برای هر مدل نیکویی برازش مقدار برآورد شده به مقدار مشاهده شده متناظر آنها را اندازه‌گیری کرد تا بدین وسیله معیاری برای سنجش خوبی یک مدل به‌دست آید. در راستای رسیدن به این هدف، در این بخش به معرفی یک شاخص می‌پردازیم.

تعریف ۴ اگر $\tilde{A}$ و $\tilde{B}$ دو عدد فازی باشند، مقدار مشابهت آن‌ها را به‌صورت زیر تعریف می‌کنیم

$$I_{UI}(\tilde{A},\tilde{B})=\frac{Card(\tilde{A}\cap\tilde{B})}{Card(\tilde{A}\cup\tilde{B})} \qquad (۲)$$

که در آن (با فرض وجود مجموع یا انتگرال)

$$Card(\tilde{A})=\begin{cases}\sum_x \tilde{A}(x) & \text{گسسته } x\\ \int_x \tilde{A}(x)\,dx & \text{پیوسته } x\end{cases}$$

و اشتراک و اجتماع نیز بر اساس عملگرهای min و max هستند.[۳]

قضیه ۴ شاخص تطبیق I_{UI} دارای ویژگی‌های زیر است

۱) $۰\le I_{UI}(\tilde{A},\tilde{B})\le ۱$

۲) $\tilde{A}=\tilde{B}\Leftrightarrow I_{UI}(\tilde{A},\tilde{B})=۱$

۳) $I_{UI}(\tilde{A},\tilde{B})=I_{UI}(\tilde{B},\tilde{A})$

۴) $\tilde{A}\subseteq\tilde{B}\subseteq\tilde{C}\Longrightarrow$

$$I_{UI}(\tilde{A},\tilde{C})\le min\{I_{UI}(\tilde{A},\tilde{B}),I_{UI}(\tilde{B},\tilde{C})\}$$

[۳] U مخفف $Union$ و I مخفف $Intersection$ است.

اثبات به‌وضوح می‌توان دید که ویژگی‌های اول و سوم در مورد این شاخص برقرار است.

ویژگی دوم: طرف $(\rightarrow)$ به‌وضوح برقرار است، برای اثبات طرف $(\leftarrow)$ توجه کنید که

$$I_{UI}(\tilde{A},\tilde{B})=۱\Rightarrow Card(\tilde{A}\cap\tilde{B})=Card(\tilde{A}\cup\tilde{B})$$

بنا بر این (در حالت پیوسته)

$$\int_R[(\tilde{A}\cup\tilde{B})(x)-(\tilde{A}\cap\tilde{B})(x)]dx=۰$$

از آنجایی که $\tilde{A}\cap\tilde{B}\subseteq\tilde{A}\cup\tilde{B}$ لذا نتیجه می‌شود که $\tilde{A}\cap\tilde{B}=\tilde{A}\cup\tilde{B}$ و این یعنی $\tilde{A}=\tilde{B}$.

ویژگی چهارم

فرض کنید $\tilde{A}\subseteq\tilde{B}\subseteq\tilde{C}$ ، لذا $\tilde{A}\cup\tilde{C}=\tilde{B}\cup\tilde{C}=\tilde{C}$ و نیز $\tilde{A}\cap\tilde{C}\subseteq\tilde{B}\cap\tilde{C}$.

با استفاده از روابط بالا داریم

$$\begin{aligned}I_{UI}(\tilde{A},\tilde{C}) &= \frac{Card(\tilde{A}\cap\tilde{C})}{Card(\tilde{A}\cup\tilde{C})}\\ &\le \frac{Card(\tilde{B}\cap\tilde{C})}{Card(\tilde{A}\cup\tilde{C})}\\ &= \frac{Card(\tilde{B}\cap\tilde{C})}{Card(\tilde{B}\cup\tilde{C})}\\ &= I_{UI}(\tilde{B},\tilde{C})\end{aligned}$$

و از طرف دیگر داریم $\tilde{A}\cap\tilde{C}=\tilde{A}\cap\tilde{B}=\tilde{A}$ و $\tilde{A}\cup\tilde{B}\subseteq\tilde{A}\cup\tilde{C}$ و لذا

$$\begin{aligned}I_{UI}(\tilde{A},\tilde{C}) &= \frac{Card(\tilde{A}\cap\tilde{C})}{Card(\tilde{A}\cup\tilde{C})}\\ &= \frac{Card(\tilde{A}\cap\tilde{B})}{Card(\tilde{A}\cup\tilde{C})}\\ &\le \frac{Card(\tilde{A}\cap\tilde{B})}{Card(\tilde{A}\cup\tilde{B})}\\ &= I_{UI}(\tilde{A},\tilde{B})\end{aligned}$$

بنا بر این

$$I_{UI}(\tilde{A},\tilde{C})\le min\{I_{UI}(\tilde{A},\tilde{B}),I_{UI}(\tilde{B},\tilde{C})\}. \quad \square$$

تعریف ۵ برای مدل رگرسیون با ضرایب فازی، MCI (میانگین شاخص‌های تطبیق [۴]) به‌صورت زیر تعریف می‌شود

$$MCI = \frac{۱}{n}\sum_{i=۱}^{n} I_{UI}(\tilde{y}_i, \tilde{Y}_i)$$

واضح است که $۰ \leq MCI \leq ۱$. از بین چند مدل، هر چه مقدار MCI برای یک مدل به یک نزدیک‌تر باشد، می‌توان نتیجه گرفت که مدل مورد نظر بهتر به مجموعه داده‌ها برازش شده است. نیز، مقدار نزدیک به صفر این معیار برای یک مدل، نشان دهنده برازش ضعیف آن مدل به داده‌ها است.

۵ مثال عددی

در این بخش مثالی را که چوی و باکلی [۸] ارائه کرده و از آن استفاده نموده‌اند، مورد توجه و بررسی قرار می‌دهیم. مجموعه داده‌های استفاده شده توسط آنها در جدول (۱) درج شده است. در این مجموعه داده‌ها، مقادیر مشاهده شده متغیر خروجی، به‌صورت اعداد فازی متقارن مثلثی هستند.

چوی و باکلی ادعا کرده‌اند که در این مجموعه داده‌ها، یک داده پرت (داده شماره ۳) وجود دارد و نشان داده‌اند که مدل رگرسیونی کمترین قدرمطلق خطای فازی (به شیوه‌ای که در مقدمه به آن اشاره شد) در مورد این داده‌ها بهتر از روش رگرسیونی کمترین مربعات خطای فازی عمل می‌کند. استفاده از مدل چوی و باکلی برای مدل‌سازی داده‌های جدول (۱) منجر به مدل زیر می‌شود

$$\tilde{Y} = -۲/۸۲۷۳ \oplus (-۰/۳۸۷۸, ۰/۲۱۳۲, ۰) \otimes x_۱ \oplus -۱/۰۱۲۵ x_۲ \oplus (۰/۶۱۸۵, ۰/۰۳۶۸, ۰/۲۱۱۲) \otimes x_۳$$

از سوی دیگر، با به‌کارگیری شیوه معرفی شده در این مقاله در مدل‌سازی داده‌های جدول (۱) به مدل زیر دست می‌یابیم

$$\tilde{Y} = -۱۷/۳۳۹ \oplus (-۰/۱۸۵۵, ۰/۱۴۹۰, ۰) \otimes x_۱ \oplus -۰/۹۳۶۰ x_۲ \oplus (۱/۶۱۸۵, ۰/۰۳۶۸, ۰/۰۸۳۷) \otimes x_۳$$

جدول ۱. داده‌های مربوط به مثال.

شماره	ورودی‌های دقیق $x_۱$	$x_۲$	$x_۳$	خروجی فازی
۱	۲/۰۰	۰/۰۰	۱۵/۲۵	(۳/۵۶ و ۵/۸۳)
۲	۰/۰۰	۵/۰۰	۱۴/۱۳	(۰/۵۲ و ۰/۸۵)
۳	۱/۱۳	۱/۵۰	۱۴/۱۳	(۸/۵۰ و ۱۳/۹۳)
۴	۲/۰۰	۱/۲۵	۱۳/۶۳	(۲/۴۴ و ۴/۰۰)
۵	۲/۱۹	۳/۷۵	۱۴/۷۵	(۱/۰۱ و ۱/۶۵)
۶	۰/۲۵	۳/۵۰	۱۳/۷۵	(۰/۹۶ و ۱/۵۸)
۷	۰/۷۵	۵/۲۵	۱۵/۲۵	(۴/۹۹ و ۸/۱۸)
۸	۴/۲۵	۲/۰۰	۱۳/۵۰	(۱/۱۳ و ۱/۸۵)

مقایسه نیکویی برازش دو مدل

معیار چوی و باکلی برای مقایسه دو مدل، مجموع خطای بین مقدار مشاهده شده و مقدار برآورد شده، است که برای هر مشاهده به‌صورت تفاضل تابع عضویت آنها تعریف می‌شود یعنی،

$$error(\tilde{y}_i, \tilde{Y}_i) = \begin{cases} \dfrac{\int_{-\infty}^{\infty} |\tilde{y}_i(x) - \tilde{Y}_i(x)|dx}{\int_{-\infty}^{\infty} \tilde{y}_i(x)dx} & \int_{-\infty}^{\infty} \tilde{y}_i(x)dx > ۱ \\ \int_{-\infty}^{\infty} |\tilde{y}_i(x) - \tilde{Y}_i(x)|dx & \int_{-\infty}^{\infty} \tilde{y}_i(x)dx \leq ۱ \end{cases}$$

در جدول ۲ مقادیر خطا برای دو مدل درج شده است.

جدول ۲. مقادیر خطا برای دو مدل

شماره	روش پیشنهادی	روش چوی و باکلی
۱	۰/۷۷۷۶	۰/۴۰۸۸
۲	۰/۳۳۱۵	۱/۲۳۲۰
۳	۱/۱۱۰۱	۱/۱۹۷۰
۴	۰/۶۹۳۳	۰/۲۵۷۱
۵	۱/۶۰۳۳	۱/۷۱۷۲
۶	۰/۳۲۹۹	۱/۹۴۷۶
۷	۱/۱۹۰۸	۱/۳۶۹۱
۸	۰/۰۰۰۰	۱/۴۸۵۲
مجموع	۶/۰۳۶۵	۹/۶۱۴۱

هم‌چنان که آشکار است مجموع مقادیر خطا برای روش پیشنهادی از مجموع مقادیر خطای روش چوی و باکلی کمتر

[۴] Mean of Capability Indices

است.
از سوی دیگر می‌توان دو مدل را بر پایه شاخص تطبیق (۲) نیز مقایسه نمود. در جدول (۳) مقادیر شاخص تطبیق برای دو مدل درج شده‌اند.
واضح است که بر اساس این معیار نیز، مدل مبتنی بر روش پیشنهاد شده در مقاله، برازش بهتری به داده‌ها دارد.

جدول (۳). مقادیر شاخص تطبیق برای دو مدل

شماره	روش پیشنهادی	روش چوی و باکلی
۱	۰/۲۴۷۳	۰/۵۹۱۲
۲	۰/۶۱۰۷	۰/۲۹۶۸
۳	۰/۰۰۰۰	۰/۰۰۹۷
۴	۰/۳۳۶۹	۰/۷۴۷۶
۵	۰/۱۱۲۴	۰/۲۷۸۴
۶	۰/۷۰۲۳	۰/۱۶۰۴
۷	۰/۰۰۲۲	۰/۰۰۹۳
۸	۱/۰۰۰۰	۰/۳۱۹۹
میانگین	۰/۳۷۶۵	۰/۳۰۱۶

مراجع

[۱] اخلاقی، سعید؛ تقی‌زاده کاخکی، حسین؛ ارقامی، ناصررضا، رگرسیون حداقل مربعات فازی، اندیشه آماری ۹ (۱۳۸۳)، ۶۲–۵۳.

[۲] ماشین‌چی، ماشاا...، رگرسیون با استفاده از پایگاه اطلاعاتی مشکک، اندیشه آماری ۵ (۱۳۷۹)، ۱۹–۱۳.

[۳] طاهری، سید محمود، آمار فازی: مروری بر گذشته و چشم اندازهای آینده، اندیشه آماری ۷ (۱۳۸۱)، ۷۲–۵۸.

[4] Arabpour A.R., Tata M. (2008), *Estimating the parameters of a fuzzy linear regression model*, Iranian Journal of Fuzzy Systems, (to appear).

[5] Bargiela A., Pedrycz W., Nakashima T. (2007), *Multiple regression with fuzzy data*, Fuzzy Sets and Systems, 158: 2169-2188.

[6] Celmins A. (1987), *Least squares model fitting to fuzzy vector data*, Fuzzy Sets and Systems, 22: 260-269.

[7] Chang P.T., Lee C.H. (1994), *Fuzzy least absolute deviations regression based on the ranking of fuzzy numbers*, Proc. of IEEE World Congress on Computaional Intelligence, pp: 1365-1369.

[8] Choi S.H., Buckley J.J. (2008), *Fuzzy regression using least absolute deviation estimators*, Soft Comput., 12: 254-263.

[9] Diamond P. (1987), *Least squares fitting of several fuzzy variables*, Proc. of Second IFSA Congress, Tokyo, pp: 20-25.

[10] Dodge Y. (1987), *An introduction to statistical data analysis L1-norm based*, In: Statistical Data Analysis Based on the L1-Norm and Related Methods, pp: 1-22.

[11] D'Urso P., Santoro A. (2006),*Goodness of fit and variable selecion in the fuzzy multiple linear regression*, Fuzzy Sets and Systems, 157: 2627-2647.

[12] Guo, P., Tanaka, H. (2006), Dual models for possibilistic regression analysis, Comput. Stat. and Data Anal. 51: 253-266.

[13] Mohammadi J., Taheri S.M. (2004), *Pedomodels fitting with fuzzy least squares regression*, Iranian Journal of Fuzzy Systems, 1(2): 45-61.

[14] Nasrabadi M.M., Nasrabadi E., Nasrabadi A.R. (2005), *Fuzzy linear regression analysis a multi-objective programming approach*, Applied Math. and Comp. 163: 245-251.

[15] Tanaka H., Vejima S., Asai K. (1982), *Linear regression analysis with fuzzy model*, IEEE Trans. System Man Cybernet, 12: 903-907.

[16] Torabi H., Behboodian J. (2007), *Fuzzy least-absolute estimates in linear models*, Communications in Statistics-Theory and Methods, 36: 1935-1944.

[17] Yang M.S., Ko C.H. (1997), *On cluster-wise fuzzy regression analysis*, IEEE Trans Systems Man Cybernet. B 27(1): 1-13.

[18] Zimmermann H.J. (1996), Fuzzy Set Theory and Its Applications, Kluwer, Dodrecht, 3rd ed.