سیستم‌های فازی و کاربردها
سال سوم، شماره دوم (پائیز و زمستان ۱۳۹۹)
صص. ۱۹۵ تا ۲۱۱
شاپای چاپی: ۲۷۱۷-۴۴۰۹
شاپای الکترونیکی: ۲۷۱۷-۳۹۸۴
DOR:20.1001.1.27174409.1399.3.2.12.9

روش اسپلاین های رگرسیونی تطبیقی چندگانه $(MARS)$ با پاسخ های فازی و کاربرد آن در پزشکی اجتماعی[1]

مینا رضایی، روشنک علی محمدی، مهشید نامداری
دانشکده علوم ریاضی، دانشگاه الزهرا
دانشکده دندانپزشکی، دانشگاه علوم پزشکی شهید بهشتی
تاریخ دریافت: ۱۳۹۹/۱۱/۱۹ تاریخ پذیرش: ۱۴۰۰/۲/۶
نوع مقاله: علمی-پژوهشی

چکیده

نخست مدل رگرسیون اسپلاین تطبیقی چندگانه $(MARS)$) را در حالتی که مقادیر متغیرهای پاسخ، فازی هستند مرور می‌کنیم. سپس این مدل را بر داده های مربوط به یک بررسی در زمینهٔ پزشکی اجتماعی به کار میگیریم. این بررسی، دربارهٔ ارتباط بین آگاهی افراد درباره بیماری های سرطان و سطح اجتماعی و اقتصادی افراد است. دانش افراد درباره سرطان به صورت مقادیر زبانی (/ کلامی) بیان شده است و در قالب مجموعه های فازی صورت بندی شده اند. نتایج مدلسازی را با نتایج به دست آمده با روش کمترین توانهای دوم فازی و کمترین مربعات خطای فازی مقایسه می‌کنیم که حاکی از برتری مدل $MARS$ است. این مقایسه برپایهٔ دو ملاک رایج نیکویی برازش انجام گرفته است.

[1] این مقاله نسخهٔ اصلاح و تکمیل شده مقاله ای است که در هفتمین کنگره مشترک سیستمهای فازی و هوشمند ایران ارائه شده است.

Mathematics Subject Classification (2010): 99X99 ;99X99 , **Email:** Mi.Rezaei@student.alzahra.ac.ir.
عبارات و کلمات کلیدی:رگرسیون اسپلاین تطبیقی چندگانه (مارس)، متغیر زبانی، داده فازی، سطح اجتماعی و اقتصادی، رگرسیون کمترین توانهای دوم فازی

۱ سرآغاز

مدل های رگرسیونی الگوهایی ریاضی را با هدف بررسی ارتباط بین متغیر پاسخ (/ وابسته / خروجی) و متغیرهای تبیینی (/ مستقل / توضیحی / ورودی/ پیش بین) فراهم میکنند. رویکرد های متفاوتی در رگرسیون فازی وجود دارد. در این مقاله به معرفی روشی جدید می‌پردازیم و با مثال کاربردی این روش را توضیح می‌دهیم. $MARS$ یک روش ناپارامتری است که هیچ فرضیه زیر بنایی درباره رابطه تابعی بین متغیرهای تبیینی و در نظر نمی‌گیرد. این روش، روشی کارآمد برای داده هایی است که دامنه وسیعی دارند و همچنین داده پرت نیز در این مدل موثر است.
در بخش ۲ به مرور و بررسی کوتاه روش $MARS$ می‌پردازیم. در بخش ۳ اهداف و انگیزه استفاده از روش $MARS$ فازی را توضیح می‌دهیم. بخش ۴ روش $MARS$ برای مدل سازی داده هایی با پاسخ فازی را شرح می‌دهیم و روش ساخت مدل را در این حالت تشریح می‌کنیم. در بخش ۵ دو معیار نیکویی برازش را معرفی می‌کنیم که در ادامه از آن استفاده می‌کنیم. بخش ۶ مثالی درباره سطح اجتماعی و اقتصادی افراد و آگاهی آنان درباره بیماری سرطان است که با استفاده از روش $MARS$ ارتباط بین آن ها را مدل بندی کرده ایم و سپس این روش را با روش رگرسیون چن و دنگ و کمترین مربعات خطا مقایسه می‌کنیم و در نهایت در بخش ۸ به نتیجه گیری می‌پردازیم.

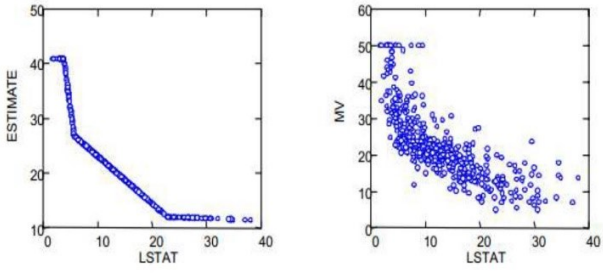
۲ رگرسیون تطبیقی چندگانه اسپلاین MARS

روش رگرسیون تطبیقی چندگانه اسپلاین (MARS) اولین بار توسط فریدمن [۷] معرفی شده و برای مسائلی با بعدهای بالا هنگامی که تعداد متغیر های پیش بین زیاد است یا حالتی که ارتباط بین متغیرها در بازه های مختلف از الگوهای مختلفی پیروی می‌کند، خوب عمل می‌کند. این روش می‌تواند به عنوان تعمیمی از رگرسیون خطی گام به گام یا اصلاحی از روش درخت رگرسیونی (CART) در نظر گرفته شود [۸]. به طوری که در روش MARS رویه پاسخ پیوسته است، اما در روش درخت رگرسیونی پیوسته نبوده و ناپیوستگی های آن در مرز نواحی افراز شده شکل می‌گیرد.
در ادامه MARS را از دید اول یعنی تعمیمی از رگرسیون خطی گام به گام مرور می‌کنیم. مدل سازی در این روش بر اساس برازش رگرسیون های خطی قطعه ای که ساده ترین نوع اسپلاین ها

است، صورت می‌گیرد. در روش MARS ، متغیرهای موثر و نقاط مرزی بازه ها (گره ها) برای هر متغیر از طریق روشی سریع تشخیص داده می‌شوند. علاوه بر جستجوی یک به یک متغیرها، MARS به جستجوی اثرات متقابل بین متغیرها تا هر مرتبه ای که مورد نظر باشد نیز می‌پردازد. مدل بهینهٔ MARS در یک فرآیند دو مرحله ای انتخاب می‌شود. در مرحلهٔ اول، MARS یک مدل بیش از حد بزرگ را با روشی خاص می‌سازد، و در مرحلهٔ دوم توابع پایه ای که کمترین سهم را در مدل دارند، حذف می‌شوند [۴].

ابتدا با شکل زیر مفهوم گره ها را معرفی می‌کنیم، در شکل ۱ قاب سمت راست نشان دهندهٔ مجموعه ای از داده ها و قاب سمت چپ یک برآورد MARS را با سه گره برای این داده ها نشان می‌دهد.

در MARS گره ها به وسیلهٔ یک روش جستجو تعیین می‌گردد. در این راستا، چنانچه معیار



شکل ۱: مجموعه ای از داده ها و یک برآورد $MARS$

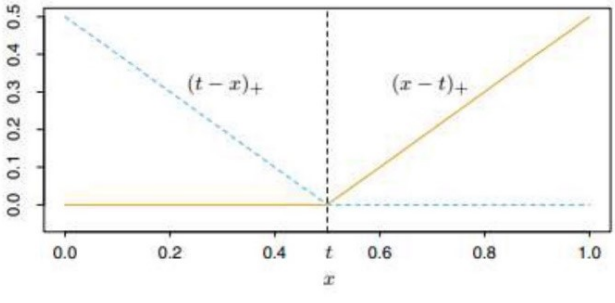
بهینگی برازش مدل رگرسیون، ضریب تعیین (R^2) باشد در این روش MARS با بررسی تعداد زیادی از گره های بالقوه، گره ای که بیش ترین مقدار ضریب تعیین را را دارد، انتخاب می‌شود [۱۲]. گاه پیدا کردن بهترین جفت گره ها، نیاز به محاسبات زیادی دارد و همچنین یافتن بهترین مجموعهٔ گره ها، وقتی که تعداد گره های مورد نیاز مجهول باشد، دشوار است. MARS مکان و تعداد مورد نیاز مورد نیاز گره ها را با یک روش پیش رو-پس رو می‌یابد. در مرحلهٔ پیش رو، مدلی که بیش برآورد دارد با گره های زیادی تولید می‌گردد. سپس گره هایی که کمترین سهم را در برازش کلی دارند، حذف می‌شوند. بنابر این انتخاب گره در مرحلهٔ پیش رو شامل مکان های نادرست زیادی برای گره ها خواهد بود، اما این گره های نادرست، احتمالا در مرحلهٔ پس رو از مدل

حذف می‌شوند. از این رو می‌توان روش MARS را یک روش تطبیقی نامید [۶].
اساس روش MARS متکی بر توابعی قطعه ای موسوم به توابع اسپلاین است که به صورت زیر تعریف می‌شوند

$$h_1(x) = (t-x)_+ = \begin{cases} t-x & \text{اگر } t > x \\ 0 & \text{در غیر این صورت} \end{cases}$$

$$h_2(x) = (x-t)_+ = \begin{cases} t-x & \text{اگر } x > t \\ 0 & \text{در غیر این صورت} \end{cases}$$

که در آن t "گره" نامیده می‌شود و نماد "+"، نشان دهندهٔ قسمت مثبت است. در این روش فرض بر آن است که گره t می‌تواند مقادیر مشاهدات متغیر x را اختیار کند، یعنی $t \in \{x_1, x_2, ..., x_n\}$. در شکل ۲، نمودار توابع $(x-0/5)_+$ و $(0/5-x)_+$ به عنوان مثال نشان داده شده است. در این شکل نمودار توابع $(x-t)_+$، خط توپر و $(t-x)_+$، خط نقطه چین می‌باشد.



شکل ۲: نمودار توابع پایه

۱.۲ روش ساختن $MARS$

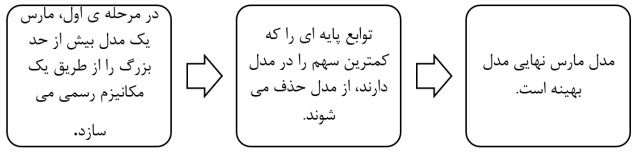
راهبرد ساختن مدل شبیه رگرسیون خطی گام به گام است. اما به جای استفاده از متغیر های ورودی، از توابع در مجموعه C و حاصل ضرب هایشان (اثرهای متقابل) استفاده می‌شود.

$$C = \{(X_j - t)_+, (t - X_j)_+\}, \quad t \in \{x_{۱j}, x_{۲j}, ..., x_{nj}\}, \quad j = ۱, ..., p.$$

اگر داده تکراری در مجموعه داده‌ها وجود نداشته باشد، مجموعهٔ C شامل $۲np$ تابع پایه خواهد بود. مدل مارس مورد نظر دارای ساختار تابعی زیر است

$$f(X) = \beta_۰ + \sum_{m=۱}^{M} \beta_m h_m(X), \qquad (۱)$$

که در آن هر $h_m(X)$ یک تابع در C یا حاصلضرب دو یا چند تا از این توابع است همچنین M تعداد توابع موجود در مدل است که پس از اعمال مرحلهٔ پس‌رو مشخص می‌گردد. با داشتن توابع h_m، ضرایب β_m بوسیله مینیمم کردن مجموع توان‌های دوم خطا، برآورد می‌شوند. . در نرم افزار R ، پکیج earth برای محاسبهٔ روش MARS تعریف شده است، به صورت کلی می‌توان گفت فرآیند کلی ساخت مدل مارس به شکل زیر است:



شکل ۳: فرآیند کلی ساخت مدل مارس

۳ اهداف و انگیزه استفاده از روش $MARS$ فازی

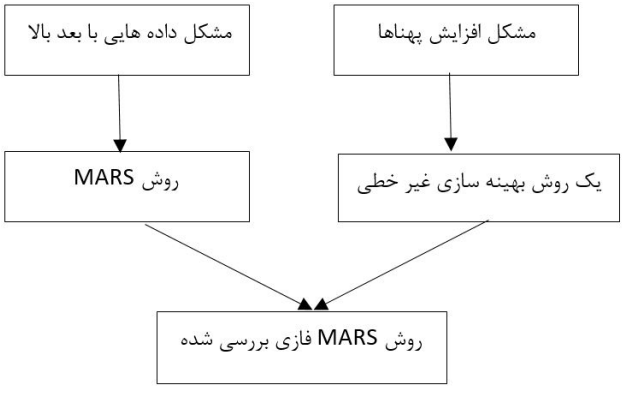
مدل های رگرسیونی (فازی یا آماری) برای بررسی ارتباط بین متغیر پاسخ و متغیرهای تبیینی الگویی را فراهم می‌کند. رگرسیون فازی را (برخلاف رگرسیون آماری) در موارد زیر می‌توان به کار برد

۱ داده ها نادقیق باشند.

۲ رابطهٔ بین متغیر ها امکانی باشد.

۳ فرضیات زیربنایی در مدل های رگرسیون آماری مورد تردید باشد.

۴ حجم نمونه کم باشد.



شکل ۴: اهداف و انگیزه استفاده از روش $MARS$ فازی

رگرسیون فازی را میتوان بسته به این که هر کدام از متغیر های تبینی، پاسخ و ضرایب مدل دقیق یا فازی در نظر گرفته شود به انواع مختلفی تقسیم بندی کرد.علاوه بر تقسیم بندی ذکر شده رگرسیون فازی رویکرد های مختلفی وجود دارد که عمده ترین آن ها عبارتند از رویکرد کمترین مربعات فازی و رویکرد رگرسیون امکانی.

رویکردهای رگرسیون فازی بسیار متنوع است. برای مروری بر آن ها می‌توان به [۱۰] [۹] [۳]

[۲] [۱] مراجعه کرد. هر یک از رویکرد ها برتری ها و ایراداتی دارند. حساسیت نسبت به داده های پرت و ابهام زیاد در مقادیر پیش بینی از جملهٔ ایرادات برخی از روش های رگرسیون فازی است. بسیاری از این رویکردها از جهات مختلف مورد بررسی قرار گرفته اند، مشکل بسیاری از روش های موجود عبارتند از، بیشتر مدل های رگرسیون امکانی و مدل های مبتنی بر کمترین مربعات فازی نسبت به داده (های) پرت حساس هستند؛ در برخی از روش ها با افزایش مقدار متغیر مستقل، پهنای برآورد شده ی متغیر پاسخ افزایش می‌یابد، که به آن مشکل افرایش پهناها گفته می‌شود؛ در برخی از روش ها که پهناهای مشاهده شده متغیر پاسخ آن روند صعودی، نزولی و یا متغیر دارند، ممکن است برآورد هایی با خطای زیاد حاصل شود. اخیرا چاچی و همکاران [۵] یک مدل رگرسیون فازی ارائه داده اند که نسبت به داده (های) پرت استوار است، و دیگر اینکه مساله افزایش پهناها را برحسب افزایش مقادیر متغیر تبیینی و پهناهای متغیر پاسخ همزمان برطرف میکند. روش ارائه شده مبتنی بر اسپلاین های رگرسیونی است، این روش تکنیک Multivariate Adaptive Regression Spline (MARS) نام دارد، آن ها با استفاده از روش MARS داده هایی که متغیر پاسخ آن ها فازی است را تحلیل کردند. در شکل ۴ که با الهام از [۵] کشیده شده است، دلایل استفاده از روش مارس فازی شرح داده شده است. در ادامه ضمن تشریح و بررسی این مدل، کاربردی از آن را در مطالعات پزشکی معرفی و بررسی میکنیم.

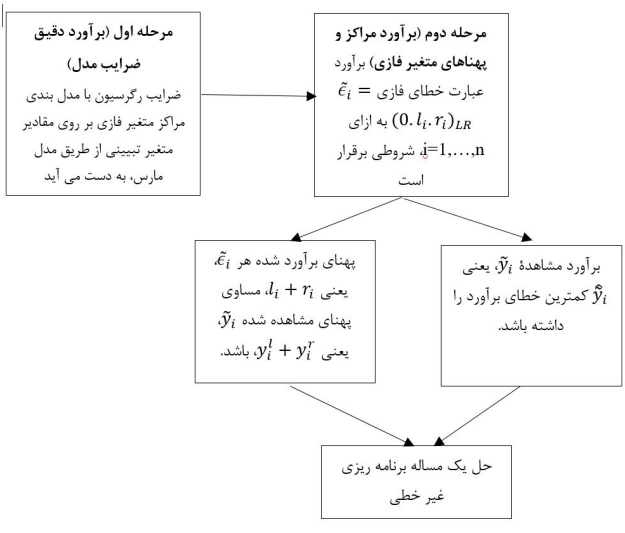
۴ روش $MARS$ با پاسخ های فازی

فرض کنید مجموعه ای از مشاهدات مربوط به یک یا چند متغیر پیش بین و یک متغیر پاسخ در اختیار داریم که مشاهدات متغیر (های) پیش بین دقیق هستند ولی مشاهدات متغیر پاسخ نادقیق (/ فازی) هستند. برای مدل بندی چنین داده هایی یک مدل رگرسیون فازی با پهناهای متغیر ارائه شده است [۵] ، مدل مربوطه به صورت زیر است

$$\tilde{y} = \beta_{\circ} + \sum_{m=1}^{M} \beta_m B_m(x_i) \oplus \tilde{\epsilon}_i \qquad i = 1, ..., n \tag{۲}$$

که در آن $\tilde{y} = (y_i, y_i^l, y_i^r)$ ، i امین مشاهده فازی متغیر پاسخ، $x_i = [x_{0i}, x_{1i}, ..., x_{ki}]$ ، i امین مشاهده حقیقی مقدار متغیرهای تبیینی، β_m ، $m = 0, 1, \ldots, M$ ، ضرایب دقیق مدل مارس برای توابع پایه[۲] $B_m(x_i)$ و $\tilde{\epsilon}_i = (0, l_i, r_i)$ خطای فازی مدل برای مشاهده i ام است.

برای برآورد پارامترها و جملات خطای فازی مدل یک روش دو مرحله ای معرفی



شکل ۵: فرایند برآورد یک مدل مارس فازی

می‌شود. در مرحله اول ضرایب دقیق $\beta = (\beta_0, \beta_1, \ldots, \beta_M)^t$ برآورد می‌شوند، و در مرحله دوم جملات خطای فازی بر اساس یک مساله بهینه سازی به دست می‌آیند.

مرحله ۱.(برآورد ضرایب دقیق مدل) ضرایب رگرسیون $\beta = [\beta_0 + \beta_1 + \ldots + \beta_M]^t$ با مدل بندی مراکز متغیر پاسخ

[2] basic function

فازی بر روی مقادیر متغیر تبیینی، از طریق مدل مارس

$$y_i = \beta_\circ + \sum_{m=1}^{M} \beta_m B_m(x_i) + \epsilon_i \qquad i = 1, ..n$$

به دست آورده می‌شود، ضرایب مدل در روش مارس با اجرای برنامه ی $earth$ و یا $MARS$ در نرم افزار R برآورد می‌شوند. سپس مراکز متغیر پاسخ فازی به صورت زیر به دست می‌آید.

مرحله ۲.(برآورد مراکز و پهناهای متغیر پاسخ فازی) عبارت خطای فازی $\tilde{\epsilon_i} = (\circ, l_i, r_i)_{LR}$ به ازای $i = 1, ..., n$ به گونه ای برآورد می‌شود که:

الف)پهنای برآورد شده هر ϵ_i یعنی $l_i + r_i$ ،مساوی پهنای مشاهده شده $\tilde{y}_i$ ، یعنی $y_i^l + y_i^r$ باشد،

ب)برآورد مشاهده $\tilde{y}_i$ یعنی $\hat{\tilde{y}}$ ،کمترین خطای برآورد را داشته باشد.

برای این منظور خطاهای فازی را از طریق حل یک مساله برنامه‌ریزی غیرخطی و با در نظر گرفتن شرط اول به عنوان قیود آن و شرط دوم به عنوان تابع هدف آن به دست آورده می‌شود که در آن تابع هدف مجموع اندازه های تشابه بین توابع عضویت مشاهده شده و برآورد شده متغیر پاسخ به صورت

$$\sum_{i=1}^{n} \frac{\int min\{\tilde{y}_i(x), \hat{\tilde{y}}_i(x)\}dx}{\int max\{\tilde{y}_i(x), \hat{\tilde{y}}_i(x)\}dx}$$

است و قیود به گونه ای اختیار می‌شوند که i امین پهنای مشاهده شده آن باشد. یعنی شرط $l_i + r_i = y_i^l + y_i^r$ برای هر i، $i = 1, ..., n$، برقرار باشد. این قید به این دلیل در نظر گرفته می‌شود که پهناهای برآورد شده کنترل شود یا به عبارتی پهناهای خیلی کم یا خیلی زیاد برای متغیر پاسخ برآورد نشود . برای در نظر گرفتن این قیود فرض می‌شود y_m^l و y_m^r به ترتیب کمترین مقدار پهنای چپ و راست مشاهدات متغیر وابسته باشند، یعنی

$$y_m^l = min\{y_1^l, ..., y_n^l\}, y_m^r = min\{y_1^r, ..., y_n^r\}$$

قسمتی از پهنای برآورد شده متغیر پاسخ مقدار ثابت $y_m^l + y_m^r$ اختیار می‌شود. واضح است که

$$D_i = (y_i^l + y_i^r) - (y_m^l + y_m^r) \geq ۰ \qquad i = ۱, ..., n$$

اما این مقدار ثابت باید از دو طرف گسترش یابد تا در نهایت مساوی با i امین پهنای مشاهده شده متغیر وابسته شود. برای این منظور، مقدار D_i به دو قسمت d_i و $D_i - d_i$ تقسیم می‌شود، که در آن $۰ \leq d_i \leq D_i$ بنابر این، قیود مساله بهینه سازی غیرخطی به صورت

$$\tilde{\epsilon}_i = (۰, y_m^l + d_i, y_m^r + D_i - d_i)_{LR} \qquad ۰ \leq d_i \leq D_i, i = ۱, ..., n$$

در نظر گرفته می‌شود، سرانجام با توجه به تابع هدف و قیود بیان شده مساله بهینه سازی غیرخطی که از طریق آن مقادیر $d_۱, ..., dn$ برآورد می‌شوند به صورت

$$max_{d_۱,...,d_n} \sum_{i=۱}^{n} \frac{\int min\{\tilde{y}_i(x), \hat{\tilde{y}}_i(x)\}dx}{\int max\{\tilde{y}_i(x), \hat{\tilde{y}}_i(x)\}dx} \hat{\tilde{y}} = \hat{\beta}_۰ + \sum_{m=۱}^{M} \hat{\beta_m} B_m(x_i) \oplus \tilde{\epsilon_i},$$

$$\tilde{\epsilon_i} = (۰, y_m^l + d_i, y_m^r + D_i - d_i)_{LR} \qquad ۰ \leq d_i \leq D_i, i = ۱, ..., n$$

در نظر گرفته می‌شود، پس از تعیین مقادیر $\hat{d}_۱, ..., \hat{d}_n$، مدل رگرسیون مارس فازی به صورت زیر حاصل می‌شود

$$\begin{aligned} \hat{\tilde{y}} &= \hat{\beta}_۰ + \sum_{m=۱}^{M} \hat{\beta_m} B_m(x_i) \oplus \tilde{\epsilon_i} \\ &= \hat{\beta}_۰ + \sum_{m=۱}^{M} \hat{\beta_m} B_m(x_i) \oplus (۰, y_m^l + d_i, y_m^r + D_i - d_i)_{LR} \qquad ۰ \leq d_i \leq D_i, i = ۱, ..., n \end{aligned}$$

به صورت کلی در شکل ۵میتوان فرایند کلی برآورد مدل مارس فازی را مشاهده کرد.

۵ نیکویی برازش

برای ارزیابی مدل های رگرسیون فازی چند ملاک توسط مولفان معرفی شده است. در این بررسی، از دو ملاک متداول برای نیکویی برازش رگرسیون فازی، به شرح زیر، استفاده می‌شود [۱۱] [۶].

$$MSM = \frac{1}{n}\sum_{i=1}^{n}\frac{\int min\{\tilde{y}_i(x), \hat{\tilde{y}}_i(x)\}dx}{\int max\{\tilde{y}_i(x), \hat{\tilde{y}}_i(x)\}dx}, \tag{۳}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\int |\tilde{y}_i - \hat{\tilde{y}}_i|dx \tag{۴}$$

توجه کنید که MSM بین صفر و یک است ولی MAE می‌تواند هر مقدار بزرگتر از صفر باشد. ضمن اینکه، MSM به بیانی ملاک شباهت است و MAE ملاکی از فاصله بین دو عدد فازی است. در مقایسهٔ بین دو مدل رگرسیون فازی هر مدلی که مقدار MAE کوچکتری (نزدیک به صفر) و یا مقدار MSM بزرگتری (نزدیک به یک داشته باشد)، برازش بهتری به داده ها دارد.

۶ کاربرد در پزشکی اجتماعی

رشته پزشکی اجتماعی یک شاخه تخصصی از پزشکی عمومی است که در آن به بررسی ارتباط بین سطح اجتماعی، اقتصادی، شرایط زندگی افراد و سلامت آن‌ها و سطح آگاهی درباره بیماری‌ها و عملکرد پزشکی آن‌ها می‌پردازد. در این علم به اثر نابرابری‌های اجتماعی بر روی بیماری‌هایی که فرد ممکن است درگیر آن شود می‌پردازد. در این مثال به بررسی آگاهی افراد از علائم و ریسک فاکتورهای سرطان پرداخته ایم و اثر متغیری به نام SES [۳] را بر سطح آگاهی مردم بررسی می‌کنیم.

[3] Social Economic Status

جدول ۱: نمونه ای از اطلاعات جمع آوری شده توسط پرسشنامه

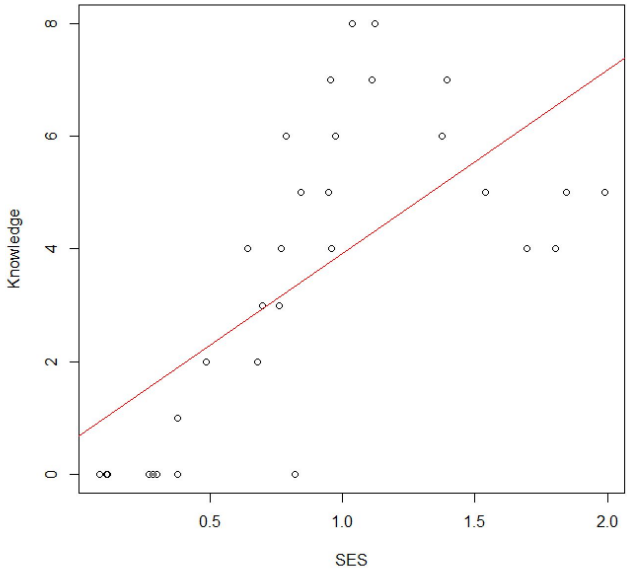
i	x_i	$\tilde{y}_i = (y_i, s_i)_T$	d_i
۱	۰/۰۸۱۹۱۳۴	$(۰،۰/۸۵)_T$	۰/۳۲
۲	۰/۰۸۱۹۱۳۴	$(۰،۰/۳۴)_T$	۰/۱۹
۳	۰/۱۰۷۴۴۶۴	$(۰،۰/۰۶)_T$	۰/۰۱
⋮	⋮	⋮	⋮
۳۱	۱/۸۴۵۰۸۳	$(۵،۰/۲۲)_T$	۰/۱۱
۳۲	۱/۹۸۹۳۵	$(۵،۰/۰۷)_T$	۰/۰۲

SES یک معیار اقتصادی و اجتماعی است که تمام مقادیر اقتصادی و اجتماعی فرد و خانواده آن فرد را مورد اندازه‌گیری قرار می‌دهد. این معیار بر اساس درآمد، سطح تحصیلات و بهره‌مندی از مجموعه‌ای از امکانات رفاهی و شغل فرد می‌باشد. همچنین محل زندگی، داشتن انواع لوازم زندگی، مساحت ملک مسکونی، استیجاری یا صاحب ملک بودن و غیره نیز مورد بررسی قرار می‌گیرد. SES معمولا برای بیان تفاوت سطح اقتصادی افراد مورد استفاده قرار می‌گیرد. حال به بررسی ارتباط بین SES افراد و سطح آگاهی آن‌ها می‌پردازیم.

یک پرسشنامه از ۳۲ نفر (والدین یک مهد کودک در تهران) پرسیده شده است، این پرسشنامه در سال ۱۳۹۵-۱۳۹۶ از افراد بالغ (۲۵-۸۰ سال) تکمیل شده است، در این پرسشنامه ۸ سوال درباره نشانه‌های سرطان دهان از افراد پرسیده شده است. اگر امتیاز فردی ۸ باشد، یعنی اطلاع آن شخص از نشانه‌های سرطان کامل است. برای اینکه اطلاعات به دست آمده دقیق باشد از افراد دو بار نظر سنجی صورت گرفته است که در صورت تفاوت در جواب‌ها این متغیر فازی در نظر گرفته شود و این تفاوت به عنوان پهنای عدد فازی فرض می‌شود. برای مثال برای مشاهده پنجم داریم:$x = ۰/۱۱$ و $y = (۰, ۰/۰۹)_T$ یعنی SES فرد مورد بررسی برابر ۰/۱۱ است و سطح دانش فرد حدود ۰ است. در اینجا میزان آگاهی افراد عدد فازی مثلثی در نظر گرفته می‌شود. مشاهدات در جدول ۱ آورده شده است. مقادر $\hat{d}_i$ محاسبه شده و مقدیر متغیرها را در ۱ مشاهده می‌کنید.

در شکل ۶ نمودار پراکنش SES افراد و میزان آگاهی آن‌ها درباره سرطان دهان مشاهده می‌کنیم. با توجه به این نمودار واضح است که یک خط راست دقت کافی برای مدل‌بندی به این داده‌ها را

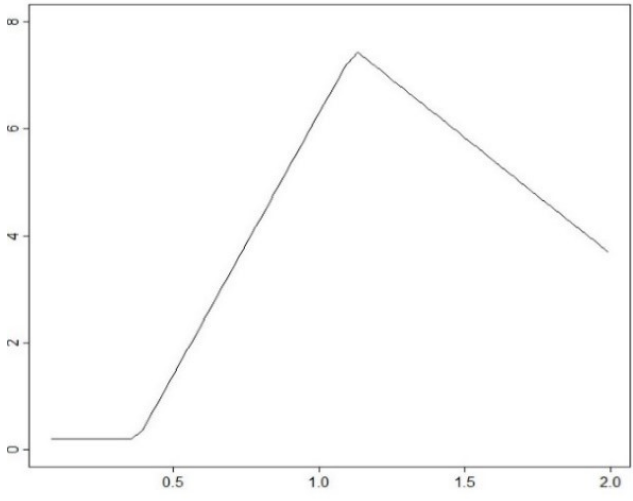
ندارد زیرا داده‌ها در بازه‌های مختلف رفتار متفاوتی از خود نشان می‌دهد. بنابراین روش مارس می‌تواند روش مناسبی برای برآورد داده‌ها باشد، می‌بینیم که مارس دارای دو گره است (در دو نقطه رفتار تابع عوض می‌شود). در ادامه با مارس فازی به برآورد متغیر پاسخ فازی $\tilde{y} = (y, s)_T$ بر اساس متغیر تبیینی x می‌پردازیم، اما برخلاف شیوه‌های متداول مانند کمترین مربعات از خطوط شکسته ای به منظور برازش یک تابع پیوسته چند ضابطه ای بهینه به داده‌ها استفاده می‌شود.



شکل ۶: نمودار پراکنش داده ها و برازش رگرسیون کمترین مربعات خطا

$$\hat{\tilde{y}}_i^{MARS} = ۰/۱۹ + ۹/۷۸max\{۰, x_i - ۰/۳۷\} - ۱۴/۱۴max\{۰, x_i - ۱/۱۲\}$$

$$\oplus\, (۰, ۰/۰۲ + \hat{d}_i, ۰/۰۲ + D_i - \hat{d}_i)_T \qquad (۵)$$



شکل ۷: مدل $MARS$ برآورد شده به داده ها

۷ مقایسه با روش های دیگر

دربخش قبل برای داده های جمع آوری شده در یک مطالعه درباره آگاهی افراد درباره بیماری افراد و سطح اجتماعی آن ها با استفاده از روش مارس مدلی را برآورد کردیم، ادامه روش ارائه شده با روش چن و دنگ [۶] و روش کمترین مربعات خطا [۱۳]

مقایسه می‌شود. مدل برآورد شده با شیوهٔ چن و دنگ به داده های پزشکی اجتماعی به صورت ۶ و با استفاده از روش کمترین قدر مطلق خطا به صورت ۷ برآورد شده است.

$$\hat{\hat{y}}_i^{CD} = ۰/۶۸ + ۳/۳۱x_i \oplus \hat{\hat{\varepsilon}}_i \qquad (۶)$$

$$\hat{\hat{y}}_i^{LA} = (۱/۱۴, ۰/۴۱)_T \oplus (۰/۱۰۷, ۰/۰۰)_T x_i \qquad (۷)$$

در جدول ۲ مقادیر MAE و MSM مدل پیشنهادی، مدل چنگ و دنگ و روش کمترین مربعات خطا مقایسه شده است.

جدول ۲: مقایسه روش $MARS$ و روش چن و دنگ برحسب دو معیار نیکویی برازش

MAE	MSM	
۰/۶۵	۰/۶۱	روش $MARS$
۱/۵۸۴	۰/۵۳	روش چن و دنگ
۰/۹۸	۰/۵۸	روش کمترین مربعات خطا

طبق جدول ۲ مشاهده می‌شود مدل MARS ، مدل مناسب تری است زیرا مقدار MSM آن بیشتر و MAE آن کمتر است در نتیجه برای داده های تحت بررسی روش مارس فازی روش کارآمد تری نسبت به روش های معمول دیگر است. در روش مارس فازی مقدار MAE ۰/۶۵ است که با مقایسه با مقدار MAE به دست آمده در مدل های چنگ و دنگ (۱/۵۸۴) و روش کمترین مقدر مطلق خطا (۰/۹۸) مقدار کمتری است و طبق بخش ۵ مدل MARS برای داده های موجود عملکرد بهتری دارد.

مقدار MSM در روش های مارس، چن و دنگ و کمترین قدر مطلق خطا به ترتیب ۰/۶۱، ۰/۵۳ و ۰/۵۸ به دست آمده که حاکی از برتری مدل مارس فازی نسبت به مدل های دیگر است.

با وجود اینکه روش کمترین مربعات خطا برای داده هایی که ارتباط خطی ندارند روشی کارا می‌باشد ؟؟ ، اما در داده های موجود در پزشکی اجتماعی روش مارش عملکرد بهتری داشته است.

۸ نتیجه گیری

در این مقاله روشی دو مرحله ای برای معرفی مدل رگرسیون فازی با پهناهای متغیر توضیح داده شد. در روش MARS فازی، ، افزایش متغیر تبیینی تاثیری بر افزایش یا کاهش پهناهای برآورد شده برای متغیر پاسخ ندارد. هم چنین روش MARS فازی می‌تواند مشاهدات فازی را که در آن ها پهناهای متغیر پاسخ، روند صعودی، نزولی، ثابت یا متغیر دارند به خوبی برآورد کند. چون در این روش، مراکز متغیر پاسخ فازی بر اساس روشی که نسبت به داده های پرت استوار است، برآورد میشود، اثرات منفی و نامناسب داده های پرت در برآورد مراکز متغیر پاسخ وارد نمی‌شوند. از طرفی می‌توان گفت که مدل پیشنهادی در مقایسه با روش چن و دنگ کمتر تحت تاثیر داده (یا

داده های) پرت قرار میگیرد، که این موضوع دلیل برتری روش پیشنهاد شده در مقاله در برابر روش آن ها است.

کاربرد این روش برپایهٔ دادههایی واقعی مبتنی بر یک بررسی در حوزه پزشکی اجتماعی بررسی شد. نتایج مدل سازی حاکی از برتری مدل رگرسیون MARS فازی بود. گفتنی است که روش چن و دنگ یک روش رگرسیون کمترین توان دوم فازی با پهناهای متغیر است که در برازش، برتری قابل توجهی نسبت به دیگر روش ها دارد. لذا مقایسه با روش چن و دنگ در واقع مقایسه با بسیاری از روش های رایج در رگرسیون فازی است.

مراجع

[۱] چاچی. ج. (۱۳۹۱) روش های آماری بر اساس اطلاعات نادقیق، رساله دکترای آمار، دانشگاه صنعتی اصفهان، دانشکده علوم ریاضی.

[۲] طاهری، س.م. (۱۳۹۶) رویکردهای ابتکاری در رگرسیون فازی، اندیشه آماری، ۲۲ (شماره ۲)، ۴۳-۵۲.

[۳] میرزایی یگانه، ش.، ارقامی، ن. ر. (۱۳۸۶) رگرسیون فازی: مروری بر چند رویکرد، اندیشهٔ آماری، ۱۲، ۳۵-۴۷.

[۴] ولی زاده، ت. (۱۳۹۱) اسپلاین های تطبیقی و غیر تطبیقی در مدل های رگرسیونی نیمه پارامتری، رساله کارشناسی ارشد آمار، دانشگاه صنعتی شاهرود، دانشکده ریاضی.

[5] Chachi J., Taheri, S. M., Pazhand, H. R., (2016) Suspended Load Estimating Using L1-Fuzzy Regression, L2-Fuzzy Regression and MARS-Fuzzy Regression Model., Hydrological Science Journal, 6, 1489-1502.

[6] Chen, S. P. and Dang, J. F. A (2008) Variable Spread Fuzzy Linear Regression Model With Higher Explanatory power and Forecasting Accuracy, Information Sicience, 38, 3973-3988.

[7] Friedman, J. H. (1990) Multivariate adaptive regression splines. Ann. Statist., 19, 11-41.

[8] Hastie, T., Tibshirani, R. and Friedman, J. H. (2009) The elements of statistical learning Data mining, Inference and prediction, 2nd Ed., Springer, New York.

[9] Khammar A., Arefi, M., Akbari, M. (2021) A general approach to fuzzy regression models based on different loss functions, Soft Computing, 25(2), 1-15.

[10] Khammar A., Arefi M., Akbari M. (2020) Robust fuzzy varying coefficient regression model based on Huber loss function, 2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS), September 2-4, 2020, Mashhad, Iran.

[11] Kelkinnama, M. and Taheri, S.M. (2012) Fuzzy Least-Absolute Regression Using Shape Preserving Operation, Information Sciences, 214, 105-120.

[12] Salford System. (2001) MARS user guide.

[13] Zeng W., Feng Q. and Li J. (2017) Fuzzy least absolute linear regression, Applied Soft Computing, 52, 1009-1019.

[14] Jiang, L. and Liao, H. (2020) Mixed fuzzy least absolute regression analysis with quantitative and probabilistic linguistic information, Fuzzy Sets and Systems, 387, 35-48.