
بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

رشته آمار، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

# رگرسیون لجستیک در محیط فازی و کاربردهای آن در علوم پزشکی

نگارنده: میترا خراباتی

استاد راهنما

دکتر محمدرضا ربیعی

استاد مشاور

دکتر فاطمه سلمانی

شهریور ۱۳۹۸

شماره:
تاریخ:

**باسمه‌تعالی**

دانشگاه صنعتی شاهرود

مدیریت تحصیلات تکمیلی

## فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم میترا خراباتی با شماره دانشجویی ۹۵۰۵۳۷۴ رشته آمار گرایش ریاضی تحت عنوان رگرسیون لجستیک در محیط فازی و کاربردها آن در علوم پزشکی که در تاریخ ۹۸/۶/۱۱ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه: عالی.....) ☑ مردود ☐

نوع تحقیق: نظری ☑ عملی ☐

| عضو هیأت داوران | نام ونام خانوادگی | مرتبهٔ علمی | امضاء |
|---|---|---|---|
| ۱ـ استادراهنمای اول | دکتر محمدرضا ربیعی | | |
| ۲- استادراهنمای دوم | | | |
| ۳- استاد مشاور | دکتر فاطمه سلمانی | | |
| ۴- نماینده تحصیلات تکمیلی | دکتر حسین باغیشنی | | |
| ۵- استاد ممتحن اول | دکتر محمد آرشی | | |
| ۶ـ استاد ممتحن دوم | دکتر احمد نزاکتی | | |

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده:

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به روح آسمانی پدرم

و وجود پر برکت مادرم که همواره مشوق من

در کسب علم و دانش بوده اند.

سپاس بی‌کران خدای راست، که دستانم را توان نوشتن از اوست تا به انجام برسانم.

در این مجال، فرصت را غنیمت شمرده، از زحمات تمامی بزرگوارانی که در این راه پشتیبانم بودند، قدردانی می‌نمایم.

وظیفه خود می‌دانم از جناب آقای دکتر محمدرضا ربیعی، استاد راهنمای ارجمندم، به خاطر تمامی محبت‌ها و راهنمایی‌های بی شائبه ایشا در تمام مراحل پایان‌نامه کمال تشکر را بنمایم.

از سرکار خانم دکتر فاطمه سلمانی، استاد مشاور گرامی که همواره با صبر و حوصله پاسخگوی سؤالات بنده بودند، متشکرم.

از داوران محترم، جناب آقای دکتر محمد آرشی و جناب آقای دکتر احمد نزاکتی و نماینده محترم تحصیلات تکمیلی، جناب آقای دکتر حسین باغیشنی بسیار متشکرم.

در انتها برخود لازم می‌دانم از خانواده خود به‌ویژه مادر بزرگوارم و دوست عزیزم خانم کتایون مسلمی، که با محبت‌های بی‌دریغشان در طی این دوران، هر چند از راه دور، بنده را مورد حمایت خودشان قرار دادند، قدردانی می‌نمایم.

از پروردگار منان خواستارم که به من توان جبران گوشه‌ای از آن همه لطف و محبت تمامی بزرگواران را ارزانی فرماید.

**میترا خراباتی**
**شهریور ۱۳۹۸**

# تعهد نامه

اینجانب **میترا خراباتی** دانشجوی کارشناسی ارشد رشته **آمار علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **رگرسیون لجستیک در محیط فازی و کاربردهای آن در علوم پزشکی** ، تحت راهنمایی **محمدرضا ربیعی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود ’’ یا “ Shahrood University of Technology ” به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

**میترا خراباتی**
**شهریور ۱۳۹۸**

## مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

# چکیده

رگرسیون لجستیک، در مدل‌سازی داده‌های آماری گسسته کمک می‌کند. مدل‌های آماری با داده‌های گسسته در علوم پزشکی (مانند مرگ/حیات، وجود بیماری/عدم بیماری و ...) کاربرد فراوان دارند. در تشخیص بیماری، پزشکان از منابع اطلاعاتی مختلفی، مانند تاریخچه بالینی بیماری، معاینات بدنی، تست‌های آزمایشگاهی و غیره استفاده می‌کنند. اما قرار دادن افراد در دو گروه بیمار و سالم اصولاً در هاله‌ای از ابهام است و مشاهدات مبهم در تشخیص بالینی، فراوان دیده می‌شود. در چنین مواردی به دلیل نادقیق بودن متغیر پاسخ، رگرسیون لجستیک مناسب نیست. برای مدل‌ساز مشاهدات پاسخ دودویی فازی، مدل رگرسیون لجستیک فازی در این پایان‌نامه ارائه شده است. برای محاسبه امکان موفقیت در مدل لجستیک فازی، اصطلاح ترم‌های زبانی مانند ...، کم، متوسط، زیاد، ... تعریف می‌شود. سپس، با استفاده از اصل گسترش، تبدیل لجیت «نسبت بخت امکانی» براساس مجموعه‌ای از مشاهدات متغیرهای تبیینی دقیق مدل‌سازی می‌شود. همچنین برای برآورد ضرایب مدل پیشنهادی از روش امکانی و روش کمترین توان دوم فازی استفاده می‌کنیم. همچنین برای ارزیابی مدل، دو روش انتخاب مدل ارائه می‌دهیم.

کلمات کلیدی: رگرسیون فازی، رگرسیون لجستیک، رگرسیون لجستیک فازی، کمترین توان دوم خطا، انتخاب مدل

# لیست مقالات مستخرج از پایان‌نامه

۱. خراباتی،میترا. ربیعی،محمدرضا،(۱۳۹۸). شیوه ای جدید در تحلیل رگرسیون لجستیک فازی، **نهمین سمینار آمار و احتمال فازی،بابلسر**

۲. ربیعی،محمدرضا. خراباتی، میترا (۱۳۹۸). معرفی بسته fuzzyreg برای تحلیل برخی از روشهای رگرسیون فازی در R، **نهمین سمینار آمار و احتمال فازی،بابلسر**

# فهرست مطالب

# فهرست تصاویر

# فهرست جداول

# فصل ۱

# مقدمات و مفاهیم رگرسیون لجستیک

## ۱.۱ مقدمه

تشخیص بیماری غالباً یک فرآیند پیچیده است که تبحر و تجربه‌ی خاصی را می‌طلبد. پزشکان برای تشخیص بیماری از منابع اطلاعاتی گوناگون استفاده می‌کنند. تاریخچه‌ی بالینی بیمار، معاینات بدنی، تست‌های آزمایشگاهی، نتایج دستگاه‌های اندازه‌گیری و شرایط و علائم از پیش تعیین شده همگی معیار تشخیص بیماری قرار می‌گیرند. اما عملاً در بسیاری از مواقع، طبقه‌بندی افراد در دو گروه بیمار و سالم هم‌چنان در هاله‌ای از ابهام است. صحبت‌های ذهنی و تا حدی غلوآمیز بیمار درباره‌ی سابقه‌ی بیماری خود، اشتباه پزشک در معاینات بدنی به‌ویژه در مورد بیماری‌هایی که علائم بالینی مشترکی ندارند، نادقیق بودن مرز بین افراد سالم و بیمار در تست‌های آزمایشگاهی، ابهام در تجزیه و تحلیل نتایج حاصل از دستگاه‌های اندازه‌گیری مانند سونوگرافی و بالاخره عدم توافق کلی در تعریف علائم بیماری‌هایی که برای تشخیص آن‌ها آزمایشات کلینیکی خاصی وجود ندارند، از جمله مواردی هستند که منجر به مشاهدات نادقیق در مطالعات بالینی می‌شوند. مدل‌سازی و تحلیل این گونه نتایج با استفاده از روش‌های معمول آماری امکان‌پذیر نمی‌باشد. برای مدل‌سازی و تحلیل چنین مشاهداتی، به‌خصوص در مواردی که حجم نمونه کم است، مدل‌سازی‌های مبتنی برمنطق و مجموعه‌های فازی راه‌گشا هست.

در فصل اول، به معرفی رگرسیون لجستیک می‌پردازیم. ابتدا رگرسیون خطی تعمیم یافته

را معرفی می‌کنیم، سپس به معرفی رگرسیون لجستیک ساده و رگرسیون لجستیک چندگانه می‌پردازیم و در نهایت فاصله اطمینان و آزمون‌های مرتبط با رگرسیون لجستیک را بررسی می‌کنیم. مطالب این بخش به صورت عمده از جان نتر و همکاران [۲۶] گرفته شده‌اند.

## ۲.۱ رگرسیون

تحلیل رگرسیونی یک روش آماری قدرتمند برای مدل‌سازی رابطه‌ی بین یک متغیر وابسته و یک یا چندین متغیر مستقل است. از اهداف اصلی تحلیل رگرسیونی کشف رابطه‌ی تابعی بین متغیر وابسته و متغیرهای مستقل است که به منظور کنترل مقادیر متغیر وابسته و یا پیش بینی آن در آینده صورت می‌گیرد. مدل ارتباطی حاکم بر متغیرهای پاسخ و مجموعه متغیرهای تبینی، توسط افراد متخصص در زمینه مورد مطالعه، بر مبنای دانش یا قضاوت‌های عینی و ذهنی آن‌ها تعیین می‌شود. این مدل که می‌تواند پارامترهای نامعلوم زیادی را در بر داشته باشد، یک مدل پارامتری نامیده می‌شود. اگر در معادله رگرسیونی تنها یک متغیر تبینی وجود داشته باشد، آن را مدل رگرسیونی خطی ساده می‌نامند. در یک مدل رگرسیونی، هدف بررسی اثر متغیرهای تبینی بر روی متغیر پاسخ است. معمولا، متغیرهای تبینی را با $x$ و متغیرهای پاسخ را با $y$ نمایش می‌دهند. هم‌چنین این ضابطه می‌تواند خطی یا غیرخطی باشد. باید توجه داشته باشیم که منظور از رابطه خطی یا غیرخطی، رابطه بین $y$ و $x$ها نیست، بلکه در واقع پارامترهای رگرسیونی به‌طور خطی یا غیرخطی وارد معادله رگرسیون می‌شوند. به‌عنوان مثال، هریک از روابط زیر خطی‌اند؛هر چند بین $x$ و $y$ ممکن است خطی نباشد:

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$y = \beta_0 + \beta_1 e^x + \epsilon$$

$$y = \beta_0 + \beta_1 log x + \epsilon$$

هم‌چنین $\epsilon$ جمله خطا می‌باشد که متغیری تصادفی و غیر قابل مشاهده است.
مثال‌هایی از یک مدل رگرسیونی عبارتند از:

۱. کشف رابطه بین سن (متغیر تبینی) و وجود بیماری کرونر قلبی (متغیر پاسخ)

۲. بررسی تاثیر میزان قند و کلسیم خون (متغیر تبینی) بر مبتلا بودن فرد به بیماری دیابت (متغیر پاسخ)

۳. بررسی تاثیر سن (متغیر تبینی) بر تعدا ضربان قلب (متغیر تبینی)

در مدل رگرسیون خطی، متغیر پاسخ $Y$ یک متغیر تصادفی پیوسته است، در حالی‌که $x$ تصادفی نبوده و توسط تحلیل‌گر کنترل و با خطای قابل اغماضی اندازه‌گیری می‌شود. بنابراین به‌ازای هر مقدار ممکن $x$ برای $y$ یک توزیع احتمال وجود دارد. به‌طور کلی ، منظور از مدل

رگرسیونی این است که $E(y|x)$ را بر حسب تابعی از $x$، یا $y$ را بر حسب تابعی از $x$، به‌همراه یک جمله خطا، $\epsilon$، که امید ریاضی آن صفر می‌شود، بنویسیم. بنابراین داریم:

$$y = \beta_0 + \beta_1 x + \epsilon \ , \ E(Y|X) = \beta_0 + \beta_1 x$$

هر دو معادل‌اند.

### ۱.۲.۱ مدل رگرسیون خطی

مدل ارتباطی (ضابطه ریاضی) حاکم برمتغیر پاسخ و مجموعه متغیرهای تبیینی، توسط افراد متخصص در زمینه مورد مطالعه، برمبنای دانش یا قضاوت‌های عینی و ذهنی آن‌ها تعیین می‌شود. فرض کنید که خط رگرسیون $y$ نسبت به $x$ به‌صورت $y = \beta_0 + \beta_1 x$ باشد، آن‌گاه مدل خطی مرتبه اول را به‌صورت

$$y = \beta_0 + \beta_1 x + \epsilon$$

می‌توانیم بنویسیم. یعنی برای $x$ داده شده، مقدار مشاهده شده‌ی $y$ مربوط به آن عبارت است از: $\beta_0 + \beta_1 x$ به اضافه مقدار خطای $\epsilon$ که ممکن است هر $y$ مشخصی خارج از خط رگرسیون قرار گیرد. به‌بیان دیگر $y$ یک متغیر تصادفی غیرقابل مشاهده و $x$ یک متغیر ریاضی (غیرتصادفی) است. معادله بالا مدلی است که براساس گمان و فرض خطی بودن ارتباط بین $x$ و $y$ بیان شده‌است.
برای مدل رگرسیون خطی چهار پذیره در نظر گرفته می‌شود:

۱. خطی بودن ضابطه تابع رگرسیون $E(y|x)$

۲. ثابت بودن واریانس جمله خطا

۳. ناهمبسته بودن مولفه‌های خطا

۴. ( برحسب نیاز) نرمال بودن خطا

تخطی از هرکدام از این پذیره‌ها (به‌ویژه پذیره اول)، می‌تواند منجر به نامناسب بودن مدل مورد نظر شود.

### ۲.۲.۱ مدل رگرسیون خطی تعمیم یافته

در سال های اخیر، مدل‌های خطی تعمیم یافته $(GLM)$ به‌طور چشم‌گیری مورد استفاده قرار گرفته‌اندو متون آماری متعددی در این زمینه به رشته تحریر درآمده‌اند. قسمتی از روند رشد این دسته از مدل‌ها را می‌توان بدین شرح نوشت: مدل‌های خطی تعمیم یافته، به‌عنوان تعمیم مدل‌های خطی، توسط نلدر و ودربرن [۲۵] معرفی شد. بررسی نظری دقیق این دسته

از مدل‌ها در مک کالا و نلدر [۲۳] آمده است. میر و همکاران [۲۲] نیز منبع خوبی برای دیدگاه شهودی این مدل‌هاست.

در مدل‌های خطی تعمیم یافته، توزیع‌هایی به‌جز نرمال نیز می‌توانند به‌عنوان توزیع متغیر پاسخ در نظر گرفته شوند. البته توزیع در نظر گرفته شده باید متعلق به خانواده نمایی یا شبیه به این خانواده باشد. [۲۳] میانگین پاسخ به‌طور مستقیم مدل‌بندی نمی‌شود بلکه تبدیلی از آن با استفاده از تابع پیوند مدل‌بندی می‌گردد.

## ۳.۱ رگرسیون لجستیک

در روش مدل‌سازی، زمانی‌که مقادیر متغیر پاسخ، گسسته است دیگر نمی‌توان از روش‌های رگرسیون خطی استفاده کرد. روش رگرسیون لجستیک توانایی مدل‌سازی این گونه مسائل را برای متغیرهای تبینی پیوسته و گسسته داراست. مدل رگرسیون لجستیکی توسط کاکس برای توصیف وابستگی یک متغیر دوتایی به مجموعه‌ای از متغیرهای پیوسته معرفی شد. [۷] در کلاس مدل‌های غیرخطی دسته‌ای از مدل‌ها وجود دارند که ذاتاً خطی هستند؛ یعنی با تبدیلی بر روی متغیرها، رابطه بین آن‌ها خطی می‌شود. مدل رگرسیون لجیت یکی از این مدل‌هاست که با تبدیل لجیت، خطی می‌شود.

رگرسیون لجستیک یکی از تکنیک‌های کاربردی برای تحلیل داده‌های رده بندی شده است که از آن برای بیان رابطه بین متغیرهای تبینی با یک متغیر پاسخ از نوع دو سطحی استفاده می‌شود. به‌عنوان نمونه، اگر نتیجه آزمایشی را به‌صورت موفقیت و شکست معرفی کنیم، در این حالت متغیر پاسخ دیگر پیوسته نبوده، و به صورت رده‌بندی شده خواهد بود. در این حالت برای رده‌بندی مشاهدات جدید، می‌توان از مدل رگرسیون لجستیک استفاده کرد که عضوی از رده مدل‌های خطی تعمیم یافته محسوب می‌شوند.

مدل رگرسیون خطی تابعی پارامتریک است و به صورت زیر بیان می‌شود:

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$E(Y|X) = \beta_0 + \beta_1 x$$

$$y = E(Y|X) + \epsilon$$

که در آن $\beta_0$ و $\beta_1$ پارامترهایی هستند که به روش کمترین مجموع مربعات خطا برآورد شده و $X$ متغیر تبینی مدل با مقدار مشاهده شده $x$ و $\epsilon$ متغیر تصادفی دارای توزیع نرمال با میانگین صفر و واریانس $\sigma^2$ می‌باشد. گفته شد در رده بندی مدل‌های خطی، مدل‌هایی وجود دارند که ذاتاً خطی هستند. به‌عبارت دیگر، با بعضی تبدیل های مناسب روی متغیرهای مدل، رابطه بین آنها خطی می‌شود. مدل رگرسیون لجستیک یکی از این مدل‌هاست. از این روش برای مدل‌سازی رابطه بین متغیر پاسخ رسته‌ای و یا یک یا چند متغیر تبینی (کمی یا کیفی) استفاده می‌شود. حال مجموعه داده‌های پاسخ را در نظر بگیرید که در آن پاسخ به صورت دودویی است

$(۰,۱)$، یعنی متغیر پاسخ گسسته است. در این‌صورت به‌جای استفاده از مدل‌بندی رگرسیون ساده، از مدل‌بندی رگرسیون لجستیک استفاده می‌کنیم.

اگر رگرسیون لجستیک احتمال وقوع موفقیت $y = ۱$ را $\pi(x)$ و احتمال وقوع شکست $y = ۰$ را $۱ - \pi(x)$ در نظر بگیریم تابع لجیت (تبدیل لجیت) که نسبت موفقیت به شکست است به‌صورت زیر است:

$$\ln\left(\frac{\pi(x)}{۱-\pi(x)}\right) = \beta_۰ + \beta_۱ x$$

استفاده می‌کنیم. هدف به‌دست آوردن $E(Y = ۱|X)$ است.

$$E(Y = ۱|X = x) = \pi(x) = \frac{exp(\beta_۰ + \beta_۱ x)}{۱ + exp(\beta_۰ + \beta_۱ x)}$$

برای یک مشاهده جدید مانند $x^*$، پیشگویی حاصل از مدل برازش شده، $\hat{\pi}(x^*)$ مقداری بین صفر و یک است. اگر هدف رده‌بندی متغیر پاسخ متناظر با این مشاهده باشد، معمولاً با تعریف مقدار آستانه، رده مشاهده پیش‌گویی می‌شود. مثلاً اگر $\hat{\pi}(x^*)$ آنگاه رده برابریک و در اگر برابر با $۱ - \hat{\pi}(x^*)$ رده برابر با صفر است. برای انجام این مرحله، ابتدا باید پارامترهای مدل یعنی ضرایب $\beta_۰$ و $\beta_۱$ برآورد شوند. برای به‌دست آوردن این برآوردها یکی از شیوه کم‌ترین توان‌های دوم خطا است.

## ۱.۳.۱ میانگین تابع پاسخ زمانی در حال دودویی

فرض کنید که مدل رگرسیون خطی ساده به‌صورت زیر تعریف شود:

$$y_i = \beta_۰ + \beta_۱ x_i + \epsilon_i \qquad y_i = ۰, ۱ \tag{۱.۱}$$

وقتی خروجی متغیر پاسخ $Y_i$ دودویی است و مقادیر ۰ یا ۱ می‌گیرد. زمانی که $E[\epsilon_i] = ۰$ باشد داریم:

$$E[\epsilon_i] = \beta_۰ + \beta_۱ x_i \tag{۲.۱}$$

چون که $Y_i$ یک متغیر تصادفی برنولی است می‌توان توزیع احتمال را به‌صورت زیر نوشت

| $Y_i$ | احتمال |
|---|---|
| ۱ | $P(Y_i = ۱) = \pi_i$ |
| ۰ | $P(Y_i = ۰) = ۱ - \pi_i$ |

بنابراین، $\pi_i$ احتمال $Y_i = ۱$ و $۱ - \pi_i$ احتمال $Y_i = ۰$ است. با استفاده از تعریف امید ریاضی داریم.

$$E[Y_i] = ۱(\pi_i) + ۰(۱ - \pi_i) = \pi_i = P(Y_i = ۱) \tag{۳.۱}$$

از (۳.۱) می‌توان نتیجه گرفت

$$E[Y_i] = \beta_0 + \beta_1 X_i = \pi_i$$

زمانی که متغیر دودویی باشد و از رگرسیون ساده استفاده کنیم ۳ مشکل داریم:

۱. نرمال نبودن عناصر خطا. برای هر متغیر پاسخ دودویی ۰ و ۱ ، هر عنصر خطا به‌صورت $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ لذا که می‌تواند دو مقدار زیر را بگیرد.

$$\text{زمانی‌که } Y_i = 1: \quad \epsilon_i = 1 - \beta_0 - \beta_1 X_i$$
$$\text{زمانی‌که } Y_i = 0: \quad \epsilon_i = -\beta_0 - \beta_1 X_i$$

واضح است که، مدل رگرسیون خطی ساده ، زمانی‌که فرض می‌شود $\epsilon_i$ به‌طور نرمال توزیع شده است، مناسب نیست.

۲. واریانس خطای غیر ثابت. مشکل دیگر این است که واریانس خطا $\epsilon_i$ زمانی‌که متغیر دودویی داریم نمی‌تواند ثابت باشد. با توجه به فرمول (۲.۱)، $V^2[Y_i]$ برای مدل رگرسیون خطی ساده می‌بایست به روش زیر به‌دست آید.

$$V^2[Y_i] = E\left[(Y_i - E[Y_i])^2\right] = (1-\pi_i)^2\pi_i + (0-\pi_i)^2(1-\pi_i)$$

یا

$$V^2[Y_i] = \pi_i(1-\pi_i) = (E[Y_i])(1-E[Y_i]) \qquad (۴.۱)$$

واریانس $\epsilon_i$ مشابه واریانس $Y_i$ است زیرا $\epsilon_i = Y_i - \pi_i$ و $\pi_i$ ثابت است.

$$V^2[\epsilon_i] = \pi_i(1-\pi_i) = E[Y_i](1-E[Y_i]) \qquad (۵.۱)$$

یا

$$V^2[\epsilon_i] = (\beta_0 + \beta_1 X_i)(1-\beta_0-\beta_1 X_i) \qquad (۶.۱)$$

در فرمول ۶.۱، $V^2[\epsilon_i]$ وابسته به $X$ است بنابراین واریانس خطا در سطوح مختلف $X$، متفاوت خواهد بود و از مجموع مربعات معمولی قابل محاسبه نیست.

۳. محدودیت بر روی تابع پاسخ. چون تابع پاسخ به‌صورت احتمالی بیان می‌شود، زمانی‌که $Y$ها دودویی هستند در این‌صورت متغیر پاسخ می‌بایست به‌صورت زیر محدود شود:

$$0 \le E[Y] = \pi \le 1$$

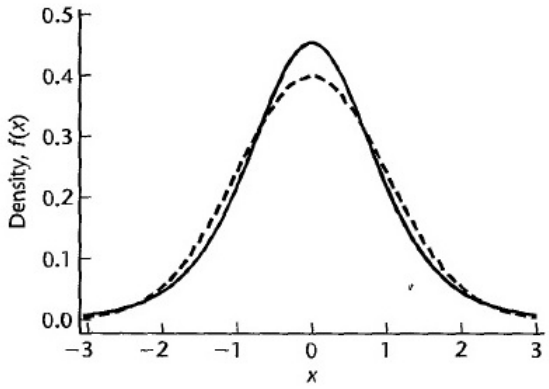
توابع پاسخ زیادی به‌طور خودکار این محدودیت را دارا نیستند. یک تابع پاسخ خطی، به‌طور مثال، ممکن است محدودیت‌های این ویژگی را در محدوده متغیر پیشگو که هدف مدل است، برطرف کند. کمترین توان دوم معمولی بهینه نمی‌شود.

## ۲.۳.۱ تابع پاسخ لجستیک

فرض کنید که مدل دارای خطای نرمال است، برای متغیر پاسخ اصلی در فرمول زیر

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (۷.۱)$$

که فرض می‌شود $\epsilon_i$ توزیع نرمال با میانگین ۰ و واریانس $\sigma_c^2$ دارد، فرض بر این است که $\phi$، دارای تابع تجمعی نرمال استاندارد برای مدل $\pi_i$ است. یک تابع خطا جایگزین، که به جای توزیع نرمال، توزیع لجستیک است. شکل (۱.۱) چگالی نرمال استاندارد و تابع چگالی لجستیک را نمایش می‌دهد، هر دو دارای میانگین ۰ و واریانس ۱ هستند. طبق شکل دُم تابع لجستیک سنگین‌تر به‌نظر می‌رسد.




شکل ۱.۱: نمودار چگالی نرمال (ـ) و چگالی لجستیک (—)

تابع لجستیک با متغیر تصادفی $\epsilon_L$، دارای میانگین ۰ و واریانس $\sigma = \frac{\pi}{\sqrt{3}}$ به‌صورت زیر است.

$$f_L(\epsilon_L) = \frac{exp(\epsilon_L)}{[1 + exp(\epsilon_L)]^2}$$

تابع توزیع تجمعی نیز به‌صورت زیر است.

$$F_L(\epsilon_L) = \frac{exp(\epsilon_L)}{1 + exp(\epsilon_L)}$$

فرض کنید که $\epsilon_i$ در معادله (۷.۱) توزیع لجستیک با میانگین ۰ و انحراف معیار $\sigma$ باشد، داریم.

$$P(Y_i = 1) = P\Big(\frac{\epsilon_i}{\sigma} \leq \beta_0^* + \beta_1^* X_i\Big)$$

که $\frac{\epsilon_i}{\sigma}$ دارای توزیع لجستک با میانگین ۰ و انحراف معیار ۱ است. با ضرب طرفین نامساوی در $\frac{\pi}{\sqrt{3}}$ داریم:

$$
\begin{aligned}
P(Y_i = ۱) = \pi_i &= P\Big(\frac{\pi}{\sqrt{۳}}\frac{\epsilon_i}{\sigma} \leq \frac{\pi}{\sqrt{۳}}\beta_۰^*\frac{\pi}{\sqrt{۳}}\beta_۱^* X_i\Big) \\
&= P(\epsilon_L \leq \beta_۰ + \beta_۱ X_i) \\
&= F_L(\beta_۰ + \beta_۱ X_i) \\
&= \frac{exp(\beta_۰ + \beta_۱ X_i)}{۱ + exp(\beta_۰ + \beta_۱ X_i)}
\end{aligned}
$$

زمانی‌که $\beta_۰ = (\pi/\sqrt{۳})\beta_۰^*$ و $\beta_۱ = (\pi/\sqrt{۳})\beta_۱^*$ باشد نشان می‌دهد که پارامترهای رگرسیون لجستیک است. به‌طور خلاصه، تابع توزیع میانگین لجستیک به صورت زیر تعریف می‌شود.

$$E[Y_i] = \pi_i = F_L(\beta_۰ + \beta_۱ X_i) = \frac{exp(\beta_۰ + \beta_۱ X_i)}{۱ + exp(\beta_۰ + \beta_۱ X_i)}$$

در یک نمایش ساده جبری از تساوی بالا می‌توان نوشت

$$E[Y_i] = \pi_i = [۱ + exp(-\beta_۰ - \beta_۱ X_i)]^{-۱} \qquad (۸.۱)$$

معکوس تابع توزیع تجمعی $F_L$ به‌صورت زیر تعریف می‌شود

$$F_L^{-۱}(\pi_i) = \beta_۰ + \beta_۱ X_i = \pi_i' \qquad (۹.۱)$$

$F_L^{-۱}(\pi_i)$ را تبدیل لجیت می‌گویند.

$$F_L^{-۱}(\pi_i) = \ln\Big(\frac{\pi_i}{۱ - \pi_i}\Big) \qquad (۱۰.۱)$$

به نسبت بالا نسبت بخت[۱] گفته می‌شود.

# ۴.۱ رگرسیون لجستیک ساده

برای برآورد پارامترهای رگرسیون لجستیک، باید از روش درستنمایی ماکزیمم استفاده کنیم. این روش، یک روش مناسب برای مقابله با مشکلات مربوط به متغیر پاسخ $Y_i$ که دودویی است، می‌باشد. در ابتدا احتیاج داریم که کارکرد احتمالی مشترک نمونه‌های مشاهده شده را تعمیم دهیم. به‌جای استفاده از توزیع نرمال برای مشاهدات $Y$ همانطور که قبلاً انجام شد، اکنون باید از توزیع برنولی برای متغیر تصادفی دودویی استفاده کنیم.

## ۱.۴.۱ مدل رگرسیون لجستیک ساده

ابتدا، به بیان یک فرمول برای مدل رگرسیون لجستیک ساده نیاز داریم. گفته شد که وقتی متغیر پاسخ دودویی باشد، احتمالات $\pi$ و $۱ - \pi$ به ترتیب مقادیر ۱ و ۰ را می‌گیرند. $Y$ متغیر

[۱] *odds ratio*

تصادفی برنولی با پارامتر $E\ [Y] = \pi$ است، که می‌توانیم مدل ساده رگرسیون لجستیک را به‌صورت زیر بیان کرد.

$$Y_i = E\left[Y_i|\ X_i\right] + \epsilon_i$$

از آن‌جایی که تابع $\epsilon_i$ وابسته به تابع برنولی پاسخ $Y_i$ است.

$$E\left[Y_i|\ X_i\right] = \pi_i = \frac{exp(\beta_\circ + \beta_\text{۱} X_i)}{\text{۱} + exp(\beta_\circ + \beta_\text{۱} X_i)} \tag{۱۱.۱}$$

فرض کنید که مشاهدات $x$ ثابت هستند. از سوی دیگر، اگر $x$ تصادفی باشد، $E\left[Y_i\right]$ به‌عنوان یک معادله شرطی با توجه به مقدار $x$ مشاهده می‌شود.

## ۲.۴.۱ تابع درستنمایی

از آنجایی که مشاهده $Y_i$ یک متغیر تصادفی برنولی عادی است داریم

$$P(Y_i = \text{۱}) = \pi_i \tag{۱۲.۱}$$

$$P(Y_i = \circ) = \text{۱} - \pi_i \tag{۱۳.۱}$$

و یا به طور خلاصه به صورت تابع احتمال بنویسیم

$$f_i(Y_i) = \pi_i^{Y_i}(\text{۱} - \pi)^{\text{۱}-Y_i} \qquad Y_i = \circ, \text{۱}; \quad i = \text{۱}, \cdots, n$$

باید توجه داشت که $f_i(\text{۱}) = \pi_i$ و $f_i(\circ) = \text{۱} - \pi_i$ است، از این رو $f(y_i)$ نشان دهنده احتمال $y_i = \circ$ یا $y_i = \text{۱}$ است.

از آن‌جایی که $Y_i$ مستقل است، تابع احتمال توام به صورت زیر نوشته می‌شود:

$$g(Y_\text{۱}, \cdots, Y_n) = \prod_{i=\text{۱}}^{n} f_i(Y_i) = \prod_{i=\text{۱}}^{n} \pi_i^{Y_i}(\text{۱} - \pi)^{\text{۱}-Y_i}$$

با لگاریتم گرفتن از تابع احتمال به راحتی می‌توان برآورد درستنمایی ماکزیمم را محاسبه کرد.

$$\begin{aligned} \ln g(Y_i, \cdots, Y_n) &= Ln \prod_{i=\text{۱}}^{n} \pi_i^{Y_i}(\text{۱} - \pi)^{\text{۱}-Y_i} \\ &= \sum_{i=\text{۱}}^{n} \left[Y_i \ln \pi_i + (\text{۱} - Y_i) \ln(\text{۱} - \pi_i)\right] \\ &= \sum_{i=\text{۱}}^{n} \left[Y_i \ln \left(\frac{\pi_i}{\text{۱} - \pi_i}\right)\right] + \sum_{i=\text{۱}}^{n} \ln(\text{۱} - \pi_i). \end{aligned} \tag{۱۴.۱}$$

از رابطه (۱۱.۱) داریم:

$$\text{۱} - \pi_i = \left[\text{۱} + exp(\beta_\circ + \beta_\text{۱} X_i)\right]^{-\text{۱}} \tag{۱۵.۱}$$

علاوه براین، از (۱۰.۱) می‌توان نتیجه گرفت.

$$Ln\left(\frac{\pi_i}{۱-\pi_i}\right)=\beta_۰+\beta_۱X_i \qquad (۱۶.۱)$$

بنابراین می‌توان با توجه به (۱۴.۱) بیان کرد

$$LnL(\beta_۰,\beta_۱)=\sum_{i=۱}^{n}Y_i(\beta_۰+\beta_۱X_i)-\sum_{i=۱}^{n}Ln\left[۱+exp(\beta_۰+\beta_۱X_i)\right] \qquad (۱۷.۱)$$

از آن‌جایی که $L(\beta_۰,\beta_۱)$ جایگزین $g(Y_۱,\cdots,Y_n)$ است می‌توانیم از آن برای برآورد پارامترهای تابع درستنمایی استفاده کنیم.

## ۳.۴.۱ برآورد پارامتر

در رگرسیون لجستیک، پارامترها معمولاً به‌روش درستنمایی ماکسیمم برآوردمی‌شوند. اما چون صورت بسته‌ای برای برآوردهای درستنمایی ماکسیمم این مدل وجود ندارد، از روش‌های عددی برای برآورد آن‌ها استفاده می‌شود.

همانند روش رگرسیون خطی می‌توان مدل رگرسیون لجستیک را به‌صورت زیر بیان کرد

$$Y=E(Y|X)+\epsilon \quad \Rightarrow \quad Y=\pi(x)+\epsilon$$

با توجه به توزیع $Y|X$، از روش درستنمایی ماکزیمم برای برآورد پارامترها استفاده می‌کنیم. از آن‌جایی که $Y|X=x$ دارای توزیع دوجمله‌ای $Y|X=x\sim B(۱,\pi(x))$ لذا داریم:

$$P(Y=y_i)=\pi(x_i)^{y_i}(۱-\pi(x_i))^{۱-y_i}$$

که در آن:

$$\begin{aligned}\pi(x_i)&=E(Y_i|X_i=x_i)=P(Y_i=۱,X_i=x_i)\\&=\frac{۱}{۱+exp(\beta_۰+\sum_{i=۱}^{n}\beta_۱x_i)}\end{aligned} \qquad (۱۸.۱)$$

در این‌صورت تابع درستنمایی براساس یک نمونه $n$تایی مستقل عبارت است از:

$$L(\beta|x)=\prod_{i=۱}^{n}P(Y=y_i)$$

که در آن $\beta=(\beta_۰,\beta_۱)^\top$ می‌باشد. با لگاریتم گرفتن از طرفین داریم:

$$\ln L(\beta|x)=\sum_{i=۱}^{n}y_i\{ln(\pi(x_i)+(۱-y_i))\}$$

با جایگذاری رابطه (۱۸.۱) داریم:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n}\{y_i \ln(\frac{1}{1+exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}) \\
&\quad + (1-y_i)\ln(\frac{exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}{1+exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)})\} \\
&= \sum_{i=1}^{n}\{\ln(\frac{exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}{1+exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}) \\
&\quad - y_i \ln(exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i))\} \\
&= \sum_{i=1}^{n}\{Ln(exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)) \\
&\quad - y_i \ln(exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)) \\
&\quad - \ln(1+exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i))\} \\
&= \sum_{i=1}^{n}\{(1-y_i)(\beta_0+\sum_{i=1}^{n}\beta_i x_i) \\
&\quad \ln(1+exp(\beta_0+\sum_{i=i}^{n}\beta_i x_i))\}
\end{aligned}$$

حال با مشتق‌گیری از رابطه بالا نسبت به پارامترها داریم:

$$\frac{d\ln L}{d\beta_0} = \sum(y_i - \pi(x_i))$$

$$\frac{d\ln L}{d\beta_1} = \sum x_i(y_i - \pi(x_i))$$

باید به این نکته توجه داشت که در مدل رگرسیون لجستیک ارتباط $Y$ و $X$ غیر خطی است و حل معادلات فوق مانند رگرسیون خطی نمی‌باشد. لذا باید برای حل معادلات فوق از روش‌های حل عددی استفاده کرد. پس از برآورد پارامترهای مورد نظر، حال می‌توان برای اختصاص دادن هر مشاهده به یکی از مقادیر ۱ یا ۰، بیشترین احتمال $P(Y_i = y_i|X_i = x_i)$ را در نظر گرفت که در آن $Y_i$ها دارای مقادیر ۰ یا ۱ می‌باشند. بنابراین $Y$ را به صفر نسبت می‌دهیم هرگاه:

$$P(Y=0|x) > P(Y=1|x)$$

$$\frac{P(Y=0|x)}{P(Y=1|x)} > 1$$

در نتیجه خواهیم داشت:

$$exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i) > 1$$

که اگر از طرفین نامساوی فوق لگاریتم بگیریم خواهد شد:

$$\beta_۰ + \sum_{i=۱}^{n} \beta_i x_i > ۰$$

$$Ln\frac{۱-\pi}{\pi} = \beta_۰ + \sum_{i=۱}^{n} \beta_i x_i > ۰$$

بنابراین قاعده تصمیم برای پیش‌بینی کلاس مورد نظر براساس روش رگرسیون لجستیک به‌صورت زیر می‌باشد:

اگر $Y = ۱$ آنگاه $\hat{\beta}_۰ + \sum \hat{\beta}_i x_i < ۰$

اگر $Y = ۰$ آنگاه $\hat{\beta}_۰ + \sum \hat{\beta}_i x_i > ۰$

همان‌طور که گفته شد روش‌های عددی برای حل این مسائل و به‌دست آوردن مقادیر $b_۰$ و $b_۱$ مناسب است. برای محاسبه، چندین بار روش‌های عددی را تکرار می‌کنیم، یکی از روش‌ها به‌طور منظم وزن دادن به کم‌ترین مربعات است. با تکیه بر برنامه‌های آماری معتبر که به‌طور مشخص برای رگرسیون لجستیک طراحی شده‌اند می‌توان مقادیر $b_۰$ و $b_۱$ را برآورد کرد.

همان‌طور که گفته شد روش‌های عددی برای حل این مسائل و به‌دست آوردن مقادیر $b_۰$ و $b_۱$ مناسب است. برای محاسبه، چندین بار روش‌های عددی را تکرار می‌کنیم، یکی از روش‌ها به‌طور منظم وزن دادن به کمترین توان های دوم است. با تکیه بر برنامه‌های آماری معتبر که به‌طور مشخص برای رگرسیون لجستیک طراحی شده‌اند می‌توان مقادیر $b_۰$ و $b_۱$ را برآورد کرد.

یک راه دیگر برای به‌دست آوردن برآورد درستنمایی ماکزیمم، باید از $\pi_۱$ برای نشان دادن مقدار برازش شده در $i$امین مقدار استفاده کنیم.

$$\hat{\pi_i} = \frac{exp(b_۰ + ۱X_i)}{۱ + exp(b_۰ + ۱X_i)} \qquad (۱۹.۱)$$

تابع پاسخ لجستیک برازش شده به صورت زیر است.

$$\hat{\pi} = \frac{exp(b_۰ + ۱X)}{۱ + exp(b_۰ + ۱X)} \qquad (۲۰.۱)$$

اگر از تعریف رگرسیون لجستیک در معادله (۹.۱) استفاده کنیم می‌توانیم با توجه به (۲۰.۱) تابع پاسخ برازش شده را به‌دست آوریم.

$$\hat{\pi}' = b_0 + b_1 X \tag{۲۱.۱}$$

در حالی‌که

$$\hat{\pi}' = Ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) \tag{۲۲.۱}$$

معادله (۲۱.۱) را تابع پاسخ لجیت برازش شده می‌نامند.

یک تابع پاسخ لجستیک برازش شده به‌دست می‌آید که معمولاً از گام‌های بعدی و امتحان کردن مناسب بودن تابع پاسخ برازش شده خوب باشد، می‌توان نتیجه‌گیری‌ها و پیشگویی‌های متنوعی داشته باشیم.

### ۴.۴.۱ تعبیر ضریب $b_1$

تعبیر ضریب $b_1$ در برآورد رگرسیونی، در تابع پاسخ لجستیک برازش شده در معادله (۲۳.۱)، تعبیری مستقیم از شیب در مدل رگرسیون خطی نیست. دلیلش این است که تاثیر افزایش یک واحد در $X$های مختلف برای مدل رگرسیون لجستیک با توجه به موقعیت شروع نقطه $X$ است. یک تعبیر برای $b_1$ را می‌توان این‌گونه گفت، به‌ازای هر واحد افزایش $X$، تابع لجستیک برازش شده که با برآورد تابع بخت، $\hat{\pi}/(1-\hat{\pi})$ شده است در $exp(b_1)$ ضرب می‌شود.

فرض کنید که مقدار تابع لجیت برازش شده (۲۱.۱) اگر $X = X_j$:

$$\hat{\pi}'(X_j) = b_0 + b_1 X_j \tag{۲۳.۱}$$

نمادگذاری $\hat{\pi}(X_j)$ نشان می‌دهد که به‌طور مشخص سطح $X$ مرتبط با مقدرا برازش شده‌است. به‌هر حال می‌توانیم فرض کنیم که مقدار تابع پاسخ لجیت برازش شده برای $X = X_j + 1$ به‌صورت زیر تعریف می‌شود:

$$\hat{\pi}'(X_j + 1) = b_0 + b_1(X_j + 1) \tag{۲۴.۱}$$

تفاضل بین دو مقدار برازش شده به صورت زیر است:

$$\hat{\pi}'(X_j + 1) - \hat{\pi}'(X_j) = b_1 \tag{۲۵.۱}$$

حال با توجه به معادله(۲۲.۱)، $\hat{\pi}(X_j)$، زمانی‌که $X = X_j$ باشد، لگاریتم برآورد $odds$ است که با $Ln(odds_1)$ مشخص می‌شود.به صورت مشابه، $\hat{\pi}(X_j + 1)$، زمانی‌که $X = X_j + 1$ باشد لگاریتم برآورد $odds$ است که با $Ln(odds_2)$ نشان داده می‌شود.بنابراین داریم:

$$\ln(odds_۲) - \ln(odds_۱) = \ln\left(\frac{odds_۲}{odds_۱}\right) = b_۱ \qquad (۲۶.۱)$$

با گرفتن آنتی‌لگ از طرفین، می‌توان مقدار نسبت $odds$ را به‌دست آورد که آن‌را نسبت بخت می‌نامند و با $\widehat{OR}$ نشان می‌دهند.

$$\widehat{OR} = \frac{odds_۲}{odds_۱} = exp(b_۱) \qquad (۲۷.۱)$$

## ۵.۱ رگرسیون لجستیک چندگانه

مدل رگرسیون لجستیک ساده که قبلا معرفی شد، یک بیان ساده برای پیش‌گویی متغیر است. در حقیقت، چندین متغیر پیش‌گو معمولاً به رگرسیون لجستیک برای به‌دست آوردن توصیف کافی و پیش‌بینی‌های مفید، نیاز دارند.

با تعمیم دادن مدل رگرسیون لجستیک ساده، به سادگی می‌توانیم $\beta_۰ + \beta_۱ X$ در (۸.۱) را با $\beta_۰ + \beta_۱ X_۱ + \cdots + \beta_(p-۱) X_(p-۱)$ جایگزینی کنیم. با ساده‌سازی فرمول‌ها، می‌توان آن را به صورت ماتریس درآورد. بردارهای زیر را در نظر بگیرید:

$$\beta_{\mathbf{p\times 1}} = \begin{bmatrix} \beta_۰ \\ \beta_۱ \\ \vdots \\ \beta_{p-۱} \end{bmatrix} \qquad \mathbf{X_{p\times 1}} = \begin{bmatrix} ۱ \\ X_۱ \\ X_۲ \\ \vdots \\ X_{p-۱} \end{bmatrix} \qquad \underset{\mathbf{p\times 1}}{\mathbf{X_i}} = \begin{bmatrix} ۱ \\ X_{i۱} \\ X_{i۲} \\ \vdots \\ X_{i.p-۱} \end{bmatrix} \qquad (۲۸.۱)$$

در این‌صورت داریم

$$X^\top \beta = \beta_۰ + \beta_۱ X_۱ + \cdots + \beta_{p-۱} X_{p-۱} \qquad (۲۹.۱)$$

$$X_i^\top \beta = \beta_۰ + \beta_۱ X_{i۱} + \cdots + \beta_{p-۱} X_{i.p-۱}$$

با تعمیم رگرسیون لجستیک ساده به رگرسیون لجستیک چندگانه داریم:

$$E[Y] = \frac{exp(X^\top \beta)}{۱ + exp(X^\top \beta)} \qquad (۳۰.۱)$$

و معادله پاسخ لجستیک (۹.۱) به صورت زیر تعمیم داده می‌شود

$$E[Y] = \left[۱ + exp(-X^\top \beta)\right]^{-۱} \qquad (۳۱.۱)$$

به‌صورت مشابه، تبدیل لجیت (۱۰.۱) به صورت

$$\pi' = ln\left(\frac{\pi}{۱-\pi}\right)$$

است. بنابراین منجر به تابع پاسخ لجیت یا پیشگوی خطی می‌شود.

$$\pi' = X^\top \beta \qquad (۳۲.۱)$$

می‌توان مدل رگرسیون لجستیک چندگانه را به‌صورت زیر بیان کرد:
$Y_i$ها متغیرهای تصادفی برنولی تصادفی با مقدار مورد انتظار $E[Y_i] = \pi_i$ هستند که

$$E[Y_i] = \pi_i = \frac{exp(X_i^\top \beta)}{۱ + exp(X_i^\top \beta)} \qquad (۳۳.۱)$$

بنابراین، مشاهدات $X$ ثابت فرض می‌شوند. با توجه به این‌که، متغیر $X$ تصادفی است $E[Y_i]$ یک میانگین که از مقادیر $X_{i۱}, \cdots, X_{i.p-۱}$ گرفته شده است.

مانند رگرسیون لجستیک ساده در (۸.۱)، رگرسیون لجستیک چندگانه در (۳۳.۱) یکنواخت و $S$شکل است. با توجه به این‌که $X\beta$ تقریبا خطی است زمانی‌که $\pi$ بین ۰٪۲ و ۰٪۸ باشد. متغیر $X$ ممکن است متفاوت از متغیر پیش‌گو باشد یا بعضاً ممکن است منحنی و یا اثر متقابل را نشان دهد. به‌هر حال، متغیرهای پیش‌گو ممکن است کمی یا کیفی باشند.

## ۱.۵.۱ برآورد پارامترهای لجستیک چندگانه

اکنون می‌خواهیم از روش درستنمایی ماکزیمم برای برآورد پارامترها در تابع لجستیک چندگانه ۳۳.۱ استفاده کنیم. تابع $log$ـ درست‌نمایی برای رگرسیون لجستیک ساده در رابطه (۱۷.۱) را می‌توان برای رگرسیون لجستیک چندگانه تعمیم داد.

$$LnL(\beta) = \sum_{i=۱}^{n} Y_i(X_i^\top \beta) - \sum_{i=۱}^{n} Ln[۱ + exp(X_i^\top \beta)] \qquad (۳۴.۱)$$

در روش‌های عددی به‌جای مقدار $\beta_۰, \cdots, \beta_{p-۱}$ از $LnL(\beta)$ استفاده شده است و به روش درستنمایی ماکزیمم $b_{n-۱}, \ldots, b_۱, b_۰$ را برآورد می‌کند. فرض کنید $b$ برداری از برآوردهای درست‌نمایی ماکزیمم باشد.

$$\mathbf{b}_{p\times ۱} = \begin{bmatrix} b_۰ \\ b_۱ \\ \vdots \\ b_{p-۱} \end{bmatrix} \qquad (۳۵.۱)$$

برازش تابع پاسخ لجستیک و برازش مقادیر را می‌توان به‌صورت زیر بیان کرد.

$$\hat{\pi} = \frac{exp(\mathbf{X}^\top \mathbf{b})}{۱ + exp(\mathbf{X}^\top \mathbf{b})} = |۱ + exp(-\mathbf{X}^\top \mathbf{b})|^{-۱} \qquad (۳۶.۱)$$

$$\hat{\pi}_i = \frac{exp(\mathbf{X}_i^\top \mathbf{b})}{۱ + exp(\mathbf{X}_i^\top \mathbf{b})} = |۱ + exp(-\mathbf{X}_i^\top \mathbf{b})|^{-۱} \quad (۳۷.۱)$$

زمانی که

$$\mathbf{X}^\top \mathbf{b} = b_۰ + b_۱ X_۱ + \cdots + b_{p-۱} X_{p-۱} \quad (۳۸.۱)$$

$$\mathbf{X}_i^\top \mathbf{b} = b_۰ + b_۱ X_{i۱} + \cdots + b_{p-۱} X_{i.p-۱} \quad (۳۹.۱)$$

# فصل ۲

# مجموعه‌ها و رگرسیون فازی

در این فصل تعاریف و قضایای نظریه مجموعه‌های فازی و رگرسیون فازی به منظور آشنایی خواننده به‌اختصار آورده شده است.

## ۱.۲ مجموعه‌های فازی

نظریه مجموعه فازی در سال ۱۹۶۵ توسط پروفسور لطفی عسکرزاده [۱] دانشمند ایرانی تبار و استاد دانشگاه برکلی آمریکا، ارائه شده است. این نظریه از زمان ارائه آن تا کنون، گسترش و تعمیق زیادی یافته و کاربردهای گوناگونی در زمینه های مختلف پیدا کرده است.

در نظریه مجموعه‌های قطعی، یک مجموعه معمولی با یک ویژگی دقیق و معین مشخص می‌شود. به‌دین ترتیب که اگر عضو مجموعه مرجع، آن خصوصیت را داشته باشد، عضو مجموعه و اگر فاقد آن خصوصیت باشد، عضو مجموعه نیست. مثلا، اول بودن اعداد حقیقی یک خصوصیت خوش تعریف است یعنی برای هر عدد حقیقی می‌توان گفت که یا عدد اول است یا عدد اول نیست. با فرض اینکه $X$ یک مجموعه مرجع باشد، برای هر زیر مجموعه معمولی از $A$ از مجموعه مرجع، که با یک خصوصیت خوش تعریف مشخص می‌شود، می‌توان یک تابع

[۱] L. A. Zadeh

نشانگر تعریف کرد که با هر عضو متعلق به $A$، مقدار صفر را اختیار کند، یعنی

$$I_A(x) = \begin{cases} ۱ & x \in A \\ ۰ & x \notin A \end{cases}$$

اما اکثر مفاهیمی که با آنها برخورد داریم و براساس آن زندگی روزمره خود را پیش می‌بریم،مفاهیمی دقیق نیستند، مانند هوای خیلی سرد، ویژگی زیبا بودن، قد خیلی بلند، مجموعه اعداد بزرگ و ... . برای این خصوصیات مرز مشخص و دقیقی را نمی‌توان تعیین کرد. برای مثال از نظر یک فرد، هوای ۵ درجه سانتی‌گراد و از نظر فرد دیگری هوای ۲ درجه سانتی‌گراد، هوای خیلی سرد است. نظریه مجموعه‌های قطعی از عهده صورت‌بندی این مفاهیم و خصوصیات برنمی‌آید.

نظریه مجموعه‌های فازی، نظریه‌ای برای صورت‌بندی و تجزیه و تحلیل این گونه مفاهیم است. این نظریه تعمیم طبیعی از نظریه مجموعه‌ای قطعی است. به‌طور مثال، فرض کنید می‌خواهیم مجموعه اعدا بزرگ را تعریف کنیم. طبیعی است که «بزرگ بودن» یک خصوصیت خوش تعریفی نیست. بنا به پیشنهاد زاده، اجازه می‌دهیم که هر $x$ در مجموعه اعداد حقیقی، به اندازه یک مقدار بین صفر و یک، عضو مجموعه اعداد بزرگ باشد، به‌طوری که هر چه آن $x$ بزرگ‌تر باشد، عدد مربوط به عضویت آن در مجموعه اعداد بزرگ به یک نزدیک‌تر، و هر چه $x$ کوچک‌تر باشد، عدد مربوط به عضویت آن در مجموعه اعداد بزرگ به صفر نزدیک‌تر است. در این‌صورت می‌گوییم این عدد به اندازه مثلا ۰/۸ عضو مجموعه اعداد بزرگ است. اساس کار نظریه مجموعه‌های فازی، گسترش مفهوم تابع نشانگر یک مجموعه، با برد $[۰,۱]$ به‌جای $\{۰,۱\}$ است.

**تعریف ۱.۱.۲.** یک زیر مجموعه فازی $A$ (از این پس کوتاهی: مجموعه فازی) از مجموعه مرجع $X$، توسط یک تابع عضویت $\mu_A(x) : X \longrightarrow [۰,۱]$ مشخص می‌شود که در آن برای هر $x \in X$، مقدار $\mu_A(x)$ میزان عضویت $x$ در مجموعه فازی $A$ است.

توجه کنید که اگر $x$ کاملا در $A$ عضو باشد، $\mu_A(x) = ۱$ و اگر اصلاً در $A$ عضو نباشد، $\mu_A(x) = ۰$ است. لذا یک مجموعه معمولی و تابع نشانگر آن حالت خاصی از مجموعه فازی و تابع عضویت آن است.

برای نشان دادن یک مجموعه فازی، روش‌های مختلفی رایج است. یک روش متداول، به کار بردن مستقیم تابع عضویت مجموعه فازی است، که از آن وقتی که مجموعه مرجع پیوسته باشد، استفاده می‌شود.

**مثال ۱.۱.۲.** فرض کنید $X = \{۱,\cdots ۷\}$. می‌خواهیم یک مجموعه فازی تعریف کنیم که اعضای آن، ویژگی نادقیق «کوچک» را داشته باشند. برای مدل‌سازی این مجموعه کافی است تابع عضویت این مجموعه فازی را مشخص کنیم. تعیین این تابع بستگی به نظر تصمیم‌گیرنده

دارد. مثلاً یک تابع می‌تواند به‌صورت زیر تعریف شود

$$A(x)=\begin{cases}۱ & x=۱\\ ۰/۹ & x=۲\\ ۰/۷ & x=۳\\ ۰/۵ & x=۴\\ ۰/۳ & x=۵\\ ۰/۱ & x=۶\\ ۰ & x=۷\end{cases}$$

برای مثال $A(۳)=۰/۷$ بدین معنی است که از نظر تصمیم‌گیر، عدد ۳ به اندازه ۰/۷ به مجموعه فازی اعداد کوچک تعلق دارد. به سخن دیگر، از نظر تصمیم‌گیر گزاره « ۳ عددی کوچک است» به‌اندازه ۰/۷ درست است.

تعریف تابع عضویت یک مجموعه فازی بستگی به نظر فرد تصمیم‌گیرنده دارد. در واقع این تعریف جنبه ذهنی و شخصی دارد. لذا می‌توان توابع عصویت مختلفی را برای یک مجموعه فازی که بیانگر یک ویژگی فازی باشد، تصور کرد.

**تعریف ۲.۱.۲.** فرض کنید $X$ یک مجموعه مرجع و $A$ یک زیر مجموعه فازی از آن باشد،

$$sup(A)=\{x\in X|A(x)>۰\}.$$

**تعریف ۳.۱.۲.** $M=sup_xA(x)$ ارتفاع یک مجموعه فازی نامیده می‌شود. اگر این عدد برابر یک باشد آن‌گاه مجموعه فازی $A$ را نرمال و در غیر این‌صورت آن را زیرنرمال می‌نامیم. البته هر مجموعه فازی زیرنرمال را می‌توان با تقسیم درجات عضویت آن بر ارتفاع $A$ نرمال کرد.

**مثال ۲.۱.۲.** در مثال قبل، $sup(A)=\{۱,۲,\cdots,۷\}$. همچنین مجموعه فازی $A$ مجموعه فازی نرمال است.

**تعریف ۴.۱.۲.** زیر مجموعه معمولی از عناصر $X$ را که درجه عضویت آنها در مجموعه $A$، حداقل به بزرگی $\alpha$ باشد، $\alpha-$ برش $A$ (مجموعه تراز $\alpha$ ام وابسته به $A$ ) گوییم و با $A_\alpha$ نشان می‌دهیم.

$$A_\alpha=\{x\in X|A(x)\geq\alpha\},\qquad ۰<\alpha\leq ۱$$

**مثال ۳.۱.۲.** فرض کنید $X=\{۰,۱,۲,\cdots\}$ مجموعه مقادیر یک متغیر تصادفی پواسن مربوط به تعداد تصادفات شبانه‌روزی در یک ناحیه باشد و فرض کنید مجموعه فازی $A$ بیانگر «تعداد

تصادفات کم» با تابع عضویت زیر تعریف شود.

$$A(x)=\begin{cases}۱ & x=۰\\ ۰/۸ & x=۱\\ ۰/۶ & x=۲\\ ۰/۴ & x=۳\\ ۰/۲ & x=۴\end{cases}$$

در این صورت چند $-\alpha$ برش $A$ عبارتند از

$$A_{۰/۲}=\{۰,۱,۲,۳,۴\} \qquad A_{۰/۵}=\{۰,۱,۲\}$$
$$A_{۰/۶}=\{۰,۱,۲\} \qquad A_{۰/۸۵}=\{۰\}$$

به‌طور خلاصه می‌توان نوشت

$$A_\alpha=\begin{cases}\{۰,۱,۲,۳,۴\} & ۰<\alpha\leq ۰/۲\\ \{۰,۱,۲,۳\} & ۰/۲<\alpha\leq ۰/۴\\ \{۰,۱,۲\} & ۰/۴<\alpha\leq ۰/۶\\ \{۰,۱\} & ۰/۶<\alpha\leq ۰/۸\\ \{۰\} & ۰/۸<\alpha\leq ۱\end{cases}$$

**قضیه ۱.۱.۲.** (اتحاد تجزیه [۲] [۳۶] ) فرض کنید $A$ مجموعه‌ای فازی از $\mathbb{R}$ با تابع عضویت $A(.)$ باشد. در این‌صورت

$$A(x)=sup_{\alpha\in[۰,۱]}\alpha I_{A_\alpha}(x)$$

که در آن $I_{A_\alpha}(.)$ تابع نشانگر $-\alpha$ برش $A$ است.

**تعریف ۵.۱.۲.** (اصل توسیع یا گسترش [۳] [۵] ) فرض کنید $X_1,\cdots,X_n$، $n$ مجموعه مرجع و $X=X_1\times\cdots\times X_n$ حاصل دکارتی آن باشد. همچنین $A_1,\cdots,A_n$، $n$ مجموعه فازی به‌ترتیب از $X_1,\cdots,X_n$ باشد. به‌علاوه $y=f(x_1,\cdots,x_n)$ یک نگاشت از $X$ به $Y$ باشد. حاصل عمل $f$ بر $n$ مجموعه فازی $A_1,\cdots,A_n$ به‌صورت مجموعه فازی $B$ از $Y$ با تابع عضویت زیر تعریف می‌شود.

$$B(y)=f(A_1,\cdots,A_n)(y)=\begin{cases}\sup\limits_{\substack{x_1,\cdots,x_n\\ y=f(x_1,\cdots,x_n)}}\min A_1(x_1),\cdots,A_n(x_n) & f^{-1}(y)\neq\emptyset\\ ۰ & f^{-1}(y)=\emptyset\end{cases}$$

[۲] Resolution Identity

[۳] Extension Priniciple

**مثال ۴.۱.۲.** فرض کنید $X_1$ و $X_2$ مجموعه اعداد حسابی، $A_1$ مجموعه فازی اعداد خیلی کوچک و $A_2$ مجموعه فازی اعداد تقریباً ۲ با توابع عضویت زیر باشند

$$A_1 = \{\frac{1}{0}, \frac{0.8}{1}, \frac{0.6}{2}, \frac{0.4}{3}, \frac{0.2}{4}\}, \quad A_2 = \{\frac{0.5}{1}, \frac{1}{2}, \frac{0.5}{3}\}$$

آنگاه براساس اصل توسیع، حاصل عمل $f(x_1, x_2) = x_1 + x_2$ بر $A_1$ و $A_2$ به‌صورت مجموعه فازی زیر به‌دست می‌آید.

$$\{\frac{0.5}{1}, \frac{1}{2}, \frac{0.8}{3}, \frac{0.6}{4}, \frac{0.5}{5}, \frac{0.4}{6}, \frac{0.2}{7}\}$$

## ۱.۱.۲ معرفی تابع fapply

برای اعمال اصل گسترش در نرم افزارهای مختلف الگوریتم‌های مختلفی وجود دارد. این امر در نرم افزار R در بسته FuzzyNumber که در پیوست آ به طور مختصر معرفی شده است، توسط تابع fapply() میسر شده است. شایان ذکر است که عدد فازی مذکور باید از نوع قطعه ای باشد. با ذکر مثال زیر، روش کار این تابع را نشان می‌دهیم. [۱۴]

**مثال ۵.۱.۲.** عدد فازی مثلثی نامتقارن $A = (0.01, 0.2, 0.18)$ را در نظر بگیرید. برای اعمال تابع $y = \ln \frac{x}{1-x}$ بر این عدد مثلثی از دستور fapply() استفاده می‌کنیم(به پیوست آ مراجعه شود). عدد مثلثی متقارن حاصل از این تبدیل $A' = (-4.59512, -1.51635, -3.89182)$ می‌باشد.

# ۲.۲ اعداد فازی

اعداد فازی، که زیر مجموعه‌های خاصی از مجموعه اعداد حقیقی هستند، در بیشتر مسایل کاربردی می‌شوند. در ادامه، مفاهیم و نتایج اصلی درباره اعداد فازی براساس مرجع [۶] ارائه می‌شود.

**تعریف ۱.۲.۲.** مجموعه فازی $N$ از $\mathbb{R}$ (اعداد حقیقی) را یک عدد فازی (حقیقی) گوییم، اگر

۱. $N$ نرمال و تک نمایی باشد.

۲. $\alpha-$برش‌های $N$، به‌ازای هر $\alpha \in (0, 1]$، به‌صورت بازه‌های بسته باشند.

منظور از تک نمایی بودن این است که تنها یک $x$ وجود دارد به‌طوری‌که $A(x) = 1$.
مجموعه همه اعداد فازی را با $\mathcal{F}(\mathbb{R})$ نشان می‌دهیم.

**مثال ۱.۲.۲.** مجموعه فازی $L$ با تابع عضویت زیر یک عدد فازی است. می‌توان $L$ را یک مدل‌سازی برای اعداد فازی تقریباً صفر تعبیر کرد.

$$L(x) = 1 - |x| \qquad -1 \leq x \leq 1$$

**تعریف ۲.۲.۲.** $N$ یک عدد دقیق [۴] (غیر فازی) با مقدار $m$ نامیده می‌شود هرگاه تابع عضویت آن به‌صورت زیر باشد

$$N(x)=\begin{cases}۱ & x=m\\ ۰ & x\neq m\end{cases}$$

مجموعه همه اعداد فازی از $\mathbb{R}$ را با $\mathcal{F}(\mathbb{R})$ نشان می‌دهیم.

**تعریف ۳.۲.۲.** عدد فازی $N$ را مثبت (منفی) گوییم و آن را با نماد $N>۰$ $(N<۰)$ نشان می‌دهیم، اگر برای هر $x\leq ۰$ $(x\geq ۰)$ ، $N(x)=۰$. عدد فازی $N$ را نامنفی (نامثبت) گوییم اگر $x<۰$ $(x>۰)$، $N(x)=۰$.

نوع خاصی از اعداد فازی، اعداد فازی LR هستند که علاوه بر اینکه ساختار ویژه‌ای دارند، برخی اعمال حسابی بر آنها از قواعد خاصی پیروی می‌کند. این ویژگی‌ها باعث شده است که در کاربردها، عمدتاً از این نوع اعداد فازی استفاده می‌شود.

**تعریف ۴.۲.۲.** (عدد فازی LR [۱۷]) اگر ساختار تابع عضویت عدد فازی $N$ به‌صورت زیر باشد

$$N(x)=\begin{cases}L\left(\frac{m-x}{\alpha}\right) & x\leq m\\ R\left(\frac{m-x}{\beta}\right) & x>m\end{cases}$$

که در آن توابع شکل $L$ و $R$، توابعی غیر صعودی از $\mathbb{R}^{+}$ به $[۰,۱]$ هستند و $L(۰)=R(۰)=۱$، آنگاه $N$ را یک عدد فازی LR نامیده و با نماد $N=(m,\alpha,\beta)_{LR}$ نشان می‌دهیم. عدد حقیقی $m$ را مقدار نما (یا مرکز یا میانه) و اعداد مثبت $\alpha$ و $\beta$ را به‌ترتیب پهنای چپ و پهنای راست $N$ می‌نامیم.
در صورتی که $L=R$ و $\alpha=\beta$، آنگاه عدد فازی را متقارن نامیده و با $N=(m,\alpha)_{LL}$ نشان می‌دهیم.
برای $\alpha=۰$ قرار می‌دهیم $L\left(\frac{m-x}{\alpha}\right)$ و برای $\beta=۰$ قرار می‌دهیم $R\left(\frac{m-x}{\beta}\right)$، لذا منظور از $X=(m,۰,۰)_{LR}$، یک عدد دقیق با مقدار $m$ است.

اعداد فازی LR متداول در تعریف زیر معرفی شده‌اند.

**تعریف ۵.۲.۲.** فرض کنید $N=(m,\alpha,\beta)_{LR}$ و $L=R$

۱. اگر $L(x)=max\{۰,۱-|x|\}$، آنگاه $N$ را یک عدد فازی مثلثی [۵] نامیده و با $N=(m,\alpha,\beta)_{T}$ نشان می‌دهیم.

۲. اگر $L(x)=e^{-x^{۲}}$، آنگاه $N$ را یک عدد فازی نرمال [۶] نامیده و با $N=(m,\alpha,\beta)_{N}$ نشان می‌دهیم.

---
[۴] crisp
[۵] Triangular
[۶] Normal

۳. اگر $L(x) = max\{۰, ۱ - x^۲\}$، آنگاه $N$ را یک عدد فازی سهموی [۷] نامیده و با $N = (m, \alpha, \beta)_P$ نشان می‌دهیم.

**تعریف ۶.۲.۲.** (بازه فازی LR ) اگر در تعریف عدد فازی شرط تک‌نمایی بودن برداشته شود، آنگاه یک بازه فازی با نماد $N = (m_۱, m_۲, \alpha, \beta)_{LR}$ و با تابع عضویت زیر خواهیم داشت

$$N(x) = \begin{cases} L\left(\frac{m_۱ - x}{\alpha}\right) & x \leq m_۱ \\ ۱ & m_۱ \leq x < m_۲ \\ R\left(\frac{x - m_۲}{\beta}\right) & x \geq m_۲ \end{cases}$$

که در آن اعداد حقیقی $m_۱$ و $m_۲$ به‌ترتیب نمای اول و نمای دوم و اعداد مثبت $\alpha$ و $\beta$ به‌ترتیب پهنای چپ و پهنای راست $N$ می‌باشند.

اگر $L(x) = R(x) = max\{۰, ۱ - |x|\}$ آنگاه بازه فازی را عدد فازی ذوزنقه‌ای [۸] نیز می‌نامند و با $N = (m_۱, m_۲, \alpha, \beta)_{T_۲}$ نشان می‌دهند.

بازه فازی می‌تواند به‌عنوان تعمیمی از عدد فازی تلقی شود. در واقع زمانی‌که $m_۱ = m_۲ = m$ آنگاه بازه فازی به عدد فازی تبدیل می‌شود.

## ۳.۲ حساب اعداد فازی

اکنون چند تعریف و قضیه را در مورد حساب اعداد فازی بیان می‌کنیم. نخست مفهوم $T-$نُرم را یادآور می‌شویم.

**تعریف ۱.۳.۲.** [۵] تابع دو متغیره $T(x, y) : I \times I \to I = [۰, ۱]$ را یک $T-$ نُرم گویند هرگاه ویژگی‌های زیر را دارا باشد.

۱. $\forall x \in [۰, ۱], T(x, ۱) = x$

۲. یکنوایی: $x_۱ \leq x_۲, y_۱ \leq y_۲ \to T(x_۱, y_۱) \leq T(x_۲, y_۲)$

۳. جابه‌جایی: $T(x, y) = T(y, x)$

۴. شرکت پذیری: $T(x, T(y, z)) = T(T(x, y), z)$

نمونه‌هایی از $T-$ نُرم‌ها در زیر آورده شده است.

$$T_m(a, b) = \min(a, b)$$

$$T_p(a, b) = ab$$

$$T_L(a, b) = \max(۰, a + b - ۱)$$

[۷] Parabolic

[۸] Trapezoidal

**گزاره ۱.۳.۲.** [۵] برای هر $T-$نُرم دلخواه داریم

$$T_w(x,y) \leq T(x,y) \leq T_m(x,y)$$

که $T_m(x,y) = min(x,y)$. همچنین

$$T_w(x,y) = \begin{cases} x & y = ۱ \\ y & x = ۱ \\ ۰ & o.w. \end{cases}$$

ضعیف‌ترین $T-$نُرم است که به نام $T-$نُرم دراستیک نیز نامیده می‌شود.

**تعریف ۲.۳.۲.** [۱۷] فرض کنید $M$ و $N$ دو عدد فازی و $* : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ یک عملگر دوتایی بر اعداد حقیقی باشند. حاصل عمل تعمیم یافته $\circledast$ بر $M$ و $N$ براساس اصل توسیع، به‌صورت یک مجموعه فازی از $\mathbb{R}$ با تابع عضویت زیر تعریف می‌شود

$$(M \circledast N)(z) = \sup_{\substack{x,y \\ x*y=z}} T(M(x), N(y))$$

که در آن $T(.,.)$ یک $T-$نُرم است.

در حالت خاص برای چهار عمل اصلی و با استفاده از $T-$نُرم مینیمم در تعریف بالا داریم

$$(M \oplus N)(z) = \sup_{z=x+y} \min[M(x), N(y)]$$

$$(M \ominus N)(z) = \sup_{z=x-y} \min[M(x), N(y)]$$

$$(M \otimes N)(z) = \sup_{z=x*y} \min[M(x), N(y)]$$

$$(M \oslash N)(z) = \sup_{z=x/y} \min[M(x), N(y)]$$

به شرط اینکه در رابطه آخر $y \neq ۰$.
از این پس، مگر در مواردی که گفته شود، همواره از عملگر مینیمم برای حساب اعداد فازی استفاده می‌کنیم.

**قضیه ۱.۳.۲.** اگر $M = (m, \alpha, \beta)_{LR}$ و $\lambda \in \mathbb{R}$ آنگاه

$$\lambda M = M\lambda = \begin{cases} (\lambda m, \lambda\alpha, \lambda\beta)_{LR}, & \lambda > ۰ \\ (\lambda m, -\lambda\beta, -\lambda\alpha)_{LR}, & \lambda < ۰ \end{cases}$$

**نتیجه ۱.۳.۲.** اگر $M = (m, \alpha, \beta)_{LR}$، آنگاه قرینه $M$ به‌صورت $(-)M = (-m, \beta, \alpha)_{LR}$ است.

**قضیه ۲.۳.۲.** اگر $M = (m, \alpha, \beta)_{LR}$ و $N = (n, \delta, \gamma)_{LR}$، آنگاه $M \oplus N$ یک عدد فازی LR به‌صورت زیر است.

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

**نتیجه ۲.۳.۲.** اگر $M = (m, \alpha, \beta)_{LR}$ و $N = (n, \gamma, \delta)_{RL}$، آنگاه $M \ominus N$ یک عدد فازی LR به‌صورت زیر است

$$M \ominus N = (m - n, \alpha + \delta, \beta + \gamma)_{LR}$$

درباره ضرب و تقسیم، رابطه دقیق مطابق بر آنچه که برای جمع و تفاضل گفته شده، وجود ندارد. اصولاً حاصل ضرب دو عدد فازی LR، یک عدد فازی LR نخواهد بود. همچنین در حالت‌هایی که حاصل تقسیم دو عدد فازی LR یک عدد فازی می‌شود، عدد حاصل از نوع LR نیست. برای ضرب و تقسیم، روابطی تقریبی پیشنهاد شده که در ادامه به بیان آنها می‌پردازیم.

## روابط تقریبی برای ضرب

الف_ اگر $M > 0$ و $N > 0$ آنگاه

$$M \otimes N = (m, \alpha, \beta)_{LR} \otimes (n, \gamma, \delta)_{LR} \simeq (mn, m\gamma + n\alpha, m\delta + n\beta)_{LR}$$

ب_ اگر $M < 0$ و $N > 0$ آنگاه

$$M \otimes N = (m, \alpha, \beta)_{LR} \otimes (n, \gamma, \delta)_{LR} \simeq (mn, n\alpha - m\delta, n\beta - m\gamma)_{LR}$$

ج_ اگر $M > 0$ و $N < 0$ آنگاه

$$M \otimes N = (m, \alpha, \beta)_{LR} \otimes (n, \gamma, \delta)_{LR} \simeq (mn, m\gamma + n\beta, m\delta + n\alpha)_{LR}$$

د_ اگر $M < 0$ و $N < 0$ آنگاه

$$M \otimes N = (m, \alpha, \beta)_{LR} \otimes (n, \gamma, \delta)_{LR} \simeq (mn, -n\beta - m\delta, -n\alpha - m\gamma)_{LR}$$

## روابط تقریبی برای تقسیم

اگر $M = (m, \alpha, \beta)_{LR}$ یک عدد فازی مثبت باشد، آنگاه

$$M^{-1} \simeq (m^{-1}, \beta m^{-2}, \alpha m^{-2})_{RL}$$

اگر $M = (m, \alpha, \beta)_{LR}$ و $N = (n, \gamma, \delta)_{LR}$ دو عدد فازی مثبت باشند، آنگاه

$$(M \otimes N) \simeq \left(\frac{m}{n}, \frac{\delta m + \alpha n}{n^2}, \frac{\gamma m + \beta n}{n^2}\right)_{LR}$$

هرچه پهناهای اعداد فازی $M$ و $N$ نسبت به مقادیر نمای آنها کوچک باشند، روابط فوق دقیق‌ترند. همچنین در همسایگی مقادیر نما، این روابط دقت بیشتری دارند.

چون در ادامه از اعداد فازی مثلثی برای ضرایب مدل رگرسیون، استفاده خواهیم کرد، لذا به یادآوری چند نکته درباره این اعداد می‌پردازیم.

می‌دانیم که هر عدد فازی مثلثی را می‌توان به‌صورت $\tilde{A} = (a, s^L, s^R)_T$ نشان داد که در آن $a$ مقدار نما (یا میانه) و $s^L$ و $s^R$ به‌ترتیب پهنای چپ و پهنای راست $\tilde{A}$ هستند. اگر $s^R \neq s^L$ آن‌گاه عدد فازی مثلثی $A$ را نامتقارن گوییم. در این حالت تابع $\tilde{A}$ را با توجه به سه مشخصه $a$، $s^L$ و $s^R$ می‌توان به‌صورت زیر نوشت

$$\tilde{A}(x) = \begin{cases} ۱ - \dfrac{a-x}{s^L} & a - s^L \leq x \leq a \\ ۱ - \dfrac{x-a}{s^R} & a < x \leq a + s^R \end{cases} \qquad (۱.۲)$$

به‌گونه‌ای دیگر نیز می‌توان این تابع عضویت را نمایش داد. یعنی پهنای راست را بر حسب پهنای چپ بیان کرد. به این‌صورت که در تابع عضویت بالا قرار دهیم $s^R = ks^L$ که در آن $k$، که عددی حقیقی و مثبت است، ضریب کشیدگی نامیده می‌شود. بنابراین عدد فازی مثلثی نامتقارن $\tilde{A}$ را می‌توان با سه‌تایی، $\tilde{A} = (a, s^L, s^R)_T$ نیز توصیف کرد.در این حالت تابع عضویت $\tilde{A}$ به صورت زیر در می‌آید.

$$\tilde{A}(x) = \begin{cases} ۱ - \dfrac{a-x}{s^L} & a - s^L \leq x \leq a \\ ۱ - \dfrac{x-a}{s^R} & a < x \leq a + ks^R \end{cases} \qquad (۲.۲)$$

اگر $s^R = s^L = s$، آن‌گاه $\tilde{A}$ را عدد مثلثی متقارن نامیده و آن را با $\tilde{A} = (a, s)_T$ نمایش می‌دهیم. در این حالت تابع عضویت $\tilde{A}$ با توجه به مشخصه $a$ و $s$ به‌صورت زیر خواهد بود.

$$\tilde{A}(x) = \begin{cases} ۱ - \dfrac{a-x}{s} & a - s \leq x \leq a \\ ۱ - \dfrac{x-a}{s} & a < x \leq a + s \end{cases} \qquad (۳.۲)$$

در رابطه (۲.۲) ،اگر $k = ۱$،آن‌گاه تابع عضویت بالا (یعنی تابع عضویت عدد فازی متقارن) به‌دست می‌آید.

**گزاره ۲.۳.۲.** فرض کنید $\tilde{A}_۱ = (a_۱, s_۱^L, s_۱^R)_T$ و $\tilde{A}_۲ = (a_۲, s_۲^L, s_۲^R)_T$ دو عدد فازی مثلثی متقارن و $c$ یک عدد حقیقی باشد،

۱. (آ) اگر $c > ۰$ آنگاه $c\tilde{A}_۱ = (ca_۱, cs_۱^L, cs_۲^R)_T$

(ب) اگر $c < ۰$ آنگاه $c\tilde{A}_۲ = (ca_۱, -cs_۲^R, -cs_۲^L)_T$

۲. $\tilde{A}_۱ \oplus \tilde{A}_۲ = (a_۱ + a_۲, s_۱^L + s_۲^L, s_۱^R + s_۲^R)_T$

## ۱.۳.۲ فاصله بین دو عدد فازی

در این بخش فرض می‌کنیم که $\tilde{A}$ و $\tilde{B}$ دو عدد فازی، و $\tilde{A}_\alpha = [a_1(\alpha), a_2(\alpha)]$ و $\tilde{B}_\alpha = [b_1(\alpha), b_2(\alpha)]$ به ترتیب $\alpha-$ برش‌های $\tilde{A}$ و $\tilde{B}$ باشند.

**تعریف ۳.۳.۲.** فاصله بین دو عدد فازی $\tilde{A}$ و $\tilde{B}$ بر پایه تابع وزنی $f(\alpha)$ به‌صورت زیر تعریف می‌شود

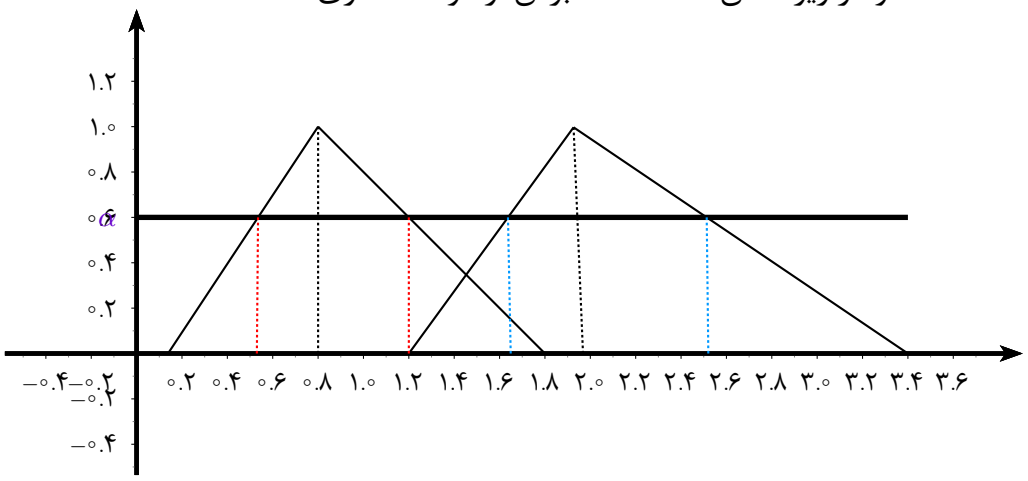
$$d(\tilde{A}, \tilde{B}) = [\int_0^1 f(\alpha) d^2(\tilde{A}_\alpha, \tilde{B}_\alpha) d_\alpha]^{\frac{1}{2}} \qquad (۴.۲)$$

که

$$d^2(\tilde{A}_\alpha, \tilde{B}_\alpha) = [a_1(\alpha) - b_1(\alpha)]^2 + [a_2(\alpha) - b_2(\alpha)]^2 \qquad (۵.۲)$$

که در آن، $f(\alpha)$ یک تابع صعودی روی بازه $[0, 1]$ است که برای آن $f(0) = 0$ و $\int_0^1 f(\alpha) d\alpha = \frac{1}{2}$. مقدار $d(\tilde{A}_\alpha, \tilde{B}_\alpha)$ فاصله بین $\alpha-$ برش‌های اعداد فازی $\tilde{A}$ و $\tilde{B}$ را اندازه می‌گیرد. تابع $f(\alpha)$ منجر می‌شود که درجات عضویت بالاتر، اهمیت بیشتری در تعیین فاصله بین دو عدد فازی $\tilde{A}$ و $\tilde{B}$ داشته باشند. شرط‌های $f(0) = 0$ و $\int_0^1 f(\alpha) d\alpha = \frac{1}{2}$ نیز این اطمینان را می‌دهد که فاصله فوق، یک تعمیم معمولی از فاصله بین دو عدد حقیقی است.

**مثال ۱.۳.۲.** نمودار زیر نشان دهنده $\alpha-$ برش از دو عدد فازی است.




طبق نمودار فوق مقادیر برای دو عدد فازی $\tilde{A}$ و $\tilde{B}$ می‌توانیم روابط زیر را بیان کنیم.

$$\tilde{A} = (a, l_A, u_A) \qquad \tilde{B} = (b, l_B, u_B) \qquad (۶.۲)$$

حال معادله خط برای $A$ را می‌نویسیم.

$$A_1 = \begin{cases} \dfrac{x - a - l_A}{l_A} & a - l_A < x < a \\ \dfrac{a + u_A - x}{u_A} & a < x < a + u_A \end{cases} \qquad (۷.۲)$$

با تلاقی خط $y=\alpha$ با دو خط بالا داریم

$$\frac{x-a+l_A}{l_A}=\alpha \Rightarrow x-\alpha l_A-l_A+a=۰ \Rightarrow x=a-(۱-\alpha)l_A \qquad (۸.۲)$$

$$\frac{a+u_A-x}{u_A}=\alpha \Rightarrow a+u_A-x-\alpha u_A \Rightarrow x=a-(۱-\alpha)u_A. \qquad (۹.۲)$$

با جایگذاری (۸.۲) و (۹.۲) در فاصله تعریف شده توسط ژو[۳۵] ۴.۲ خواهیم داشت.

$$\begin{aligned}
d(\tilde{A},\tilde{B}) &= \left[\int_۰^۱ f(\alpha)d^۲(\tilde{A},\tilde{B})d\alpha\right]^۲ \\
&= \int_۰^۱ \alpha\left(a_۱(\alpha)-b_۱(\alpha)\right)^۲ d\alpha + \int_۰^۱ \alpha\left(a_۲(\alpha)-b_۲(\alpha)\right)^۲ d\alpha \\
&= \int_۰^۱ \alpha\left[a-(۱-\alpha)l_A-b+(۱-\alpha)l_B\right]^۲ d\alpha \\
&\quad + \int_۰^۱ \alpha\left[a+(۱-\alpha)u_A-b-(۱-\alpha)u_B\right]^۲ d\alpha \\
&= \int_۰^۱ \alpha\left[(a-b)+(۱-\alpha)(l_A-l_B)\right] + \int_۰^۱ \alpha\left[(a-b)-(۱-\alpha)(u_A-u_B)\right] \\
&= \int_۰^۱ \alpha(a-b)^۲ d\alpha + \int_۰^۱ \alpha(۱-\alpha)^۲(l_B-l_A)^۲ d\alpha \\
&\quad + \int_۰^۱ ۲\alpha(۱-\alpha)(a-b)(l_B-l_A)d\alpha + \int_۰^۱ \alpha(a-b)^۲ d\alpha \\
&\quad + \int_۰^۱ \alpha(۱-\alpha)^۲(u_B-u_A)^۲ d\alpha + \int_۰^۱ ۲\alpha(۱-\alpha)(a-b)(u_B-u_A)d\alpha \\
&= (a-b)^۲ + \frac{۱}{۱۲}\left[(l_B-l_A)^۲ + (u_B-u_A)^۲\right] \\
&\quad + \frac{۱}{۳}(a-b)\left[(l_B-l_A)-(u_B-u_A)\right] \qquad (۱۰.۲)
\end{aligned}$$

## ۴.۲ رگرسیون فازی

همان‌گونه که گفته شد، مدل‌های رگرسیونی برای ایجاد ارتباط بین یک متغیر وابسته و تعدادی متغیر مستقل به‌کار می‌روند. برای ساختن این مدل‌ها، نیاز به مشاهداتی از متغیرهای مورد مطالعه داریم. در رگرسیون کلاسیک فرض می‌شود که این متغیرها و مشاهدات مربوط به آنها، دقیق هستند. همچنین به علت عدم برازش کامل داده‌ها، در رگرسیون کلاسیک، یک جمله خطای تصادفی به مدل افزوده می‌شود. درباره این جمله خطای تصادفی و توزیع احتمالی آن مفروضاتی مانند نرمال بودن، ناهمبسته بودن، ثبات واریانس و ... در نظر گرفته می‌شود، به‌گونه‌ای که بتوان بر پایه این فرض‌ها، استنباط‌های آماری مانند برآورد پارامترها، پیش‌بینی

مقدار متغیر وابسته، آزمون فرض‌ها و مواردی دیگر را انجام داد. در این‌باره، به‌کارگیری تبدیلات مختلفی بر روی مشاهدات ممکن است کمک کنند تا فرض‌های اولیه مدل برقرار باشند.

اما در بسیاری اوقات ممکن است یک یا چند فرض از فرض‌های بالا برقرار نباشد یا اینکه نتوان از درستی بعضی فرض‌ها اطمینان حاصل کرد. مسلماً در این موارد مدل‌های رایج اعتبار و کارایی لازم را ندارند. یکی از راه حل‌ها در این زمینه، رگرسیون نیرومند است که به‌ویژه زمانی که حتی با انجام تبدیلات نتوان این فرضیات را برقرار کرد، کارایی دارند.

مشکل دیگر اینکه ممکن است در بررسی، مشاهدات مربوط به یک یا چند متغیر، نادقیق باشند و یا نادقیق گزارش شده باشند. همچنین ممکن است که متغیرهای مورد مطالعه، ذاتاً دارای ارتباطی نادقیق و مبهم باشند. یکی از شیوه‌های مهم جایگزین رگرسیون در چنین مواقعی، استفاده از رگرسیون فازی است. البته زمانی که فرضیات مربوط به خطای تصادفی برقرار نباشد یا مثلاً به دلیل حجم نمونه کم نتوان از درستی آنها اطمینان حاصل کرد، نیز می‌توان از رگرسیون فازی استفاده کرد.

بعد از معرفی نظریه مجموعه‌های فازی توسط زاده، از اوایل دهه هشتاد میلادی، گسترش آن در آمار شروع شد. یکی از مهم‌ترین و کاربردی‌ترین ارتباطات بین این نظریه و آمار کلاسیک، رگرسیون فازی است.

در یک تقسیم‌بندی کلی، بسته به اینکه هر کدام از متغیرهای وابسته، مستقل و ضرایب مدل، دقیق و یا فازی در نظر گرفته شوند، انواع رگرسیون فازی را می‌توان به حالات زیر تقسیم کرد. [۵]

- رگرسیون فازی در حالتی که ارتباط بین متغیرهای دقیق، فازی فرض شود. به‌عبارت دیگر، ضرایب معادله رگرسیونی فازی در نظر گرفته شوند. این مدل‌ها به مدل‌های رگرسیون امکانی معروفند.
- رگرسیون فازی در حالتی که متغیرها و یا مشاهدات مربوط به آنها نادقیق و فازی باشند.
- مدل با ورودی و خروجی فازی و ضرایب غیر فازی.( به این مدل رگرسیون الزامی نیز گفته می‌شود.)
- مدل با خروجی و ضرایب فازی. (اکثر مقالات ارائه شده در رگرسیون فازی در ارتباط با این گونه مدل‌ها می‌باشند.)
- مدل با ورودی‌های دقیق و خروجی فازی و ضرایب دقیق به‌علاوه یک خطای فازی.
- رگرسیون در حالتی که هم متغیرها و هم ضرایب مدل فازی در نظر گرفته شوند.

علاوه بر تقسیم‌بندی ذکر شده، انواع رگرسیون فازی با درنظر گرفتن جنبه های دیگر، به‌صورت زیر تقسیم بندی می‌شوند:[۸]

ـ از نظر نوع اعداد فازی
عامل دیگری که به تنوع رگرسیون فازی می‌افزاید، نوع توابعی است که برای نشان دادن تابع عضویت کمیت‌های فازی انتخاب می‌شوند. همان‌طور که در قبل ذکر شد انواع اعداد فازی عبارت‌اند از: اعداد فازی مثلثی، اعداد فازی سهموی، اعداد فازی نرمال.
ـ از نظر انواع روش
مشابه با رگرسیون آماری، اصل یا اصولی که برآورد پارامترهای مدل براساس آن‌ها انجام می‌پذیرند نیز عاملی در ایجاد تنوع در رگرسیون فازی است. در رگرسیون فازی، دو روش مهم، رگرسیون امکانی و رگرسیون کمترین توان‌های دوم است.
ـ از نظر تعامل ضرایب
وجود یا عدم وجود تعامل بین ضرایب فازی مدل، رگرسیون فازی را به دو نوع رگرسیون فازی رگرسیون فازی با تاثیر متقابل و رگرسیون فازی بدون تاثیر متقابل تقسیم می‌کند.

## ۵.۲ رگرسیون امکانی

رگرسیون امکانی نخستین بار توسط تاناکا و همکاران [۳۳] پیشنهاد شد. آنها یک مدل برنامه‌ریزی خطی برای محاسبه ضرایب فازی مدل رگرسیونی با وروردی و خروجی دقیق ارائه کردند که تابع هدف آن که باید حداقل شود به‌صورت مجموع پهناهای ضرایب (که اعداد فازی مثلثی متقارن هستند) است و محدودیت‌های آن باعث می‌شود که به ازای مقدار مشخصی از $h$، مقدار متغیر پاسخ دست کم به اندازه $h$ در خروجی فازی برآورد شده عضو باشد. این روش سپس به مدلی با خروجی فازی نیز تعمیم یافت. در این رویکرد چون تابع عضویت مجموعه‌های فازی اغلب به‌عنوان توزیع‌های امکانی توصیف می‌شود، روش مرتبط را رگرسیون امکانی می‌نامند.

از جمله مشکلات رویکرد اولیه تاناکا این است که جواب مساله وابسته به مقیاس متغیرهای ورودی $x_j$ است. ایراد دیگر اینکه در برخی مواقع تعداد زیادی از ضرایب فازی مدل به‌صورت دقیق برآورد می‌شوند. تاناکا و همکاران [۳۰] این نقیصه را بدین صورت اصلاح کردند که به‌جای مجموع پهناهای ضرایب، مجموع پهناهای پاسخ‌های برآورد شده حداقل شوند [۱۹].

مشکل دیگر مساله تاناکا این است که اگر به‌جای $x_j$ از $(x_j - \bar{x}_j)$ (که $\bar{x}_j$ میانگین متغیر $x_j$ است)، استفاده شود، تابع برآورد شده بسیار متفاوت خواهد بود. نقیصه دیگر حساسیت نسبت به داده‌های پرت است. برای حل این مشکل، تاناکا و همکاران [۳۳] از مدل عطفی [۹] استفاده کردند که به‌جای اینکه $-\alpha$ برش‌های پاسخ‌های برآورد شده از مدل، شامل $-\alpha$ برش‌های پاسخ‌های داده شده متناظرشان باشند، فقط نیاز دارد که هر $-\alpha$ برش پاسخ مشاهده شده، اشتراک داشته باشد.

[۹] Conjunctive

محققان دیگر، تغییرات اندک یا قابل توجه در روش رگرسیون امکانی ایجاد کردند. از بین آنها، ساویک و پدریچ [۲۸] به دلیل اینکه مدل تاناکا نماهای داده‌ها را در محاسبه ضرایب رگرسیون به حساب نمی‌آورد، حال آنکه این نماها به دلیل داشتن درجه عضویت از اهمیت بیشتری برخوردارند، یک روش دو مرحله‌ای پیشنهاد کردند. ابتدا نماهای ضرایب فازی را با استفاده از رگرسیون کمترین مربعات معمولی به‌دست آوردند و سپس پهناهای آنها را با روش تاناکا و با استفاده مراکز برآورده شده، برآورد کردند. تاناکا و ایشیبوجی [۳۱] پیشنهاد کردند که نماهای ضرایب با استفاده از رگرسیون کمترین مربعات معمولی برآورد شوند اما برای به‌دست آوردن پهناهای ضرایب، از یک نوع برنامه‌ریزی درجه دویی مشابه مساله قبل استفاده می‌شود. چانگ و لی [۲۱] خاطر نشان کردند که زمانی که روند نماهای اعدا فازی در مغایرت با روند پهناها باشد، روش تاناکا اکثراً منجر به تفسیر اشتباه می‌شود. برای اجتناب از این مشکل، آنها قید مثبت بودن پهناهای ضرایب را از روش تاناکا حذف کردند.

## ۶.۲ رگرسیون کمترین توان‌های دوم فازی

رهیافت دیگری در رگرسیون فازی، روش کمترین توان‌های دوم فازی است که اولین بار توسط دیاموند [۱۲] و کلمینس [۱۰] ارئه شد. این روش تعمیمی ازرگرسیون کمترین توان‌های دوم کلاسیک است که معمولاً بر پایه تعریف‌هایی برای فاصله بین اعداد فازی است. شیوه‌های متنوعی برپایه روش کمترین توان‌های دوم فازی پیشنهاد شده است که این تنوع عوامل مختلفی دارد مانند دقیق یا فازی بودن مشاهدات مربوط به متغیرهای تبیینی، دقیق یا فازی بودن مشاهدات متغیر پاسخ، نوع تعریف فاصله بین دو عدد فازی، معیارهای نیکویی برازش و برخی موارد دیگر. نمونه‌هایی از کارهای صورت گرفته در این زمینه به شرح زیر است.

دیاموند [۱۲] با معرفی فاصله‌ای بر اعداد فازی، مدل‌هایی را برای برازش کمترین توان‌های دوم فازی برای مدل با ورودی دقیق و خروجی فازی و مدل با ورودی و خروجی فازی پیشنهاد داد. دیاموند و کورنر [۱۶] رگرسیون فازی را براساس متغیرهای تصادف فازی ارائه کردند. کلمینس [۱۰] رگرسیون کمترین توان‌های دوم را برای توابع عضویت درجه دو مطرح کرد. ژو و لی [۳۵] یک روش کمترین توان‌های دوم را برای برآورد ضرایب فازی مدل با استفاده از متری بر مبنای $\alpha-$برش‌ها ارائه کردند و با یک معیار نیکویی برازش، خوبی مدل رگرسیونی را اندازه گرفتند. محمدی و طاهری [۲۴] این روش را برای برازش مدل‌های رگرسیونی در خاک شناسی موسوم به پدوترانسفر به‌کار بردند.

کائو چیو [۲۰] یک روش دو مرحله‌ای برای برآورد مدل رگرسیون خطی فازی پیشنهاد نمودند. در مرحله اول با مقادیر غیر فازی شده مشاهدات فازی و به روش کمترین مربعات معمولی، ضرایب دقیق مدل را برآورد کردند و سپس در مرحله دوم به منظور افزایش قدرت توضیح دهندگی داده‌ها، یک عبارت خطای فازی به مدل برآورد شده اضافه کردند. کاپی و همکاران [۱۱] یک مدل رگرسیون خطی کلی را برای مطالعه وابستگی یک متغیر پاسخ فازی از

نوع LR، روی یک مجموعه متغیرهای دقیق تبیینی از طریق یک روند برآورد کمترین توان‌های دوم تکرار شونده، معرفی کردند.

## ۱.۶.۲ مدل رگرسیون خطی با ضرایب فازی

در یک تقسیم‌بندی کلی، انواع رگرسیون فازی را می‌توان به سه حالت زیر تقسیم کرد:

الف: رگرسیون فازی در حالتی که ارتباط بین متغیرها فازی فرض می‌شود. به‌عبارت دیگر ضرائب معادله رگرسیونی، فازی در نظر گرفته می‌شوند.

ب: رگرسیون فازی در حالتی که متغیرها و یا مشاهدات مربوط به متغیرها نادقیق و فازی هستند.

ج: رگرسیون فازی در حالتی که هم متغیرها و هم ضرائب مدل، فازی در نظر گرفته می‌شوند

در اینجا مدل رگرسیون خطی با ضرایب فازی (حالت الف) را مورد مطالعه قرار داده‌ایم. به مدل‌های رگرسیون با ضرایب فازی، گاهی مدل‌های رگرسیون امکانی هم گفته می‌شود. صورت کلی مدلی که مورد بحث قرار می‌گیرد عبارت است از

$$\tilde{\mathbf{Y}} = f(\mathbf{x}, A) = \tilde{A}_0 + \tilde{A}_1 x_1 + \tilde{A}_2 x_2 + \cdots + \tilde{A}_{p-1} x_{p-1} \qquad (۱۱.۲)$$

که در آن $\tilde{Y}$ متغیر وابسته یا اصطلاحاً خروجی فازی است، $\mathbf{x} = (x_1, x_2, \cdots, x_{p-1})$ بردار متغیرهای پاسخ یا اصطلاحاً بردار ورودی (با مقادیر حقیقی) و $\mathbf{A} = \{\tilde{A}_0, \tilde{A}_1, \ldots, \tilde{A}_{p-1}\}$ یک مجموعه از اعداد فازی است. مسئله‌ای که درصدد حل آن هستیم به این‌صورت مطرح می‌شود: مجموعه‌ای از داده‌های معمولی به صورت $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ در اختیار داریم. می‌خواهیم پارامترهای فازی $\tilde{A}_0, \tilde{A}_1, \ldots, \tilde{A}_{p-1}$ را به‌گونه‌ای تعیین کنیم که مدل (۱۱.۲) براساس برخی از معیارهای نیکویی برازش، بهترین برازش را به داده‌های مذکور داشته باشد.

### تابع عضویت مدل با ضرایب نامتقارن

اگر $A_j$ها $j = 0, 1, \cdots, p-1$ اعداد فازی نامتقارن و $x_i$ها نیز اعداد حقیقی و مثبت باشند، آن‌گاه بنا به رابطه ۱۱.۲ و گزاره (۲.۳.۲) ، $\tilde{Y}$ یعنی خروجی فازی نیز یک عدد فازی مثلثی نامتقارن به‌صورت $\tilde{Y} = (f^c(\mathbf{x}), f_s^L(\mathbf{x}), f_s^R(\mathbf{x}))$ که در آن $f^c(\mathbf{x})$ نما و $f_s^L(\mathbf{x})$ پهنای چپ و $f_s^R(\mathbf{x})$ پهنای راست $\tilde{Y}$ می‌باشد و به‌صورت زیر به‌دست می‌آیند.

$$f^c(\mathbf{x}) = a_0 + a_1 x_1 + \cdots + a_{p-1} x_{p-1}$$

$$f_s^L(\mathbf{x}) = s_0^L + s_1^L x_1 + \cdots + s_{p-1}^L x_{p-1}$$

$$f_s^R(\mathbf{x}) = s_0^R + s_1^R x_1 + \cdots + s_{p-1}^R x_{p-1}$$

به بیان دیگر تابع عضویت $\tilde{Y}$ عبارت است از

$$\tilde{Y}(y) = \begin{cases} ۱ - \dfrac{f^c(\mathbf{x}) - y}{f_s^L(\mathbf{x})} & f^c(\mathbf{x}) - f_s^L(\mathbf{x}) \leq y \leq f^c(\mathbf{x}) \\ ۱ - \dfrac{y - f^c(\mathbf{x})}{f_s^R(\mathbf{x})} & f^c(\mathbf{x}) < y \leq f^c(\mathbf{x}) + f_s^R(\mathbf{x}) \end{cases} \qquad (۱۲.۲)$$

اگر بخواهیم تابع عضویت بالا را برحسب ضرایب کشیدگی بیان کنیم، قرار می‌دهیم $s_i^R = k_i s_i^L$، بدین ترتیب $f_s^R(\mathbf{x})$ به‌صورت زیر تغییر می‌یابد

$$f_s^R(\mathbf{x}) = k_۰ s_۰^L + k_۱ s_۱^L x_۱ + \cdots + k_{p-۱} s_{p-۱}^L x_{p-۱}$$

# فصل ۳

# رگرسیون لجستیک فازی با رویکرد امکانی

در این فصل می‌خواهیم رگرسیون لجستیک فازی در حالتی که مقادیر متغیر پاسخ مبهم و به‌صورت اعداد فازی هستند و نشان دهنده امکان موفقیت در موقعیت می‌باشند را معرفی کنیم. در همین راستا بخت امکانی و ترم زبانی را معرفی می‌کنیم و سپس مدل را صورت‌بندی و شیوه برآورد پارامتر به روش پوراحمد و همکاران [۱] را ارائه خواهیم کرد.

## ۱.۳ معرفی مدل رگرسیون لجستیک فازی

همان‌طور که در فصل اول بیان شد، مدل رگرسیون لجستیک مبتنی بر چند فرض اساسی هستند. برخی از این مفروضات را به‌طور خلاصه بیان می‌کنیم. نتایج مربوط به متغیر وابسته به‌صورت ۰ و ۱ است، رابطه بین متغیرها دقیق است، خطای مدل یک خطای تصادفی است. همچنین برای آنکه بتوانیم برخی تحلیل‌های پارامتری آماری را در مورد مدل انجام دهیم بعضاً حجم نمونه باید به اندازه کافی بزرگ باشد.

اما در بعضی موارد در عمل مفروضات بیان شده برقرار نیستند. مانند اینکه مشاهده متغیر پاسخ، تخصیص مربوط به موفقیت یا شکست، منجر به نتایج کاملا دقیق نشود، یا اینکه ممکن است عدم اطمینان بیان شده با خطای مدل، از نوع تصادفی نباشد. علاوه بر این موارد، در

بعضی موارد به دلیل هزینه‌های بالا و ملاحظات اخلاقی در جمع‌آوری اطلاعات، اخذ تعداد زیاد نمونه غیر ممکن یا دشوار است. برای غلبه بر محدودیت‌هایی که گفته شدو برای تحلیل داده‌ها و مدل‌بندی داده‌های مبهم نظریه مجموعه‌های فازی و نظریه امکان می‌تواند کار ساز باشد.

برای مدل‌بندی رابطه بین متغیر پاسخ دودویی و متغیرهای تبیینی فرض می‌کنیم که متغیرهای تبیینی دقیق و پاسخ‌ها فازی هستند و بردار متغیرها به‌صورت $(x_{i0}, x_{i1}, \cdots, x_{in}, \tilde{Y}_i)$ باشد. از آنجایی که متغیرهای پاسخ دودویی غیر دقیق (فازی) جمع‌آوری می‌شوند، بنابراین فرض احتمالات توزیع برنولی را نمی‌توان برای این داده‌ها در نظر گرفت. لذا احتمال $(P(Y_i = 1) = \pi_i)$ را نمی‌توان محاسبه کرد. به‌عنوان راه حل به‌جای آنکه اندازه پیشامد محاسبه شود، اندازه امکان آن را می‌توان در نظر گرفت. امکان وجه دیگری از ابهام است.

**تعریف ۱.۱.۳.** [۴] فرض کنید $\mu_i, i = 1, 2, \cdots, m$ نشان دهنده امکان موفقیت باشد، آنگاه در نظر می‌گیریم:

الف یک مقدار واقعی دقیق $\mu_i$ به‌طوری‌که $\mu_i \in \mathbb{R}; 0 \leq \mu_i \leq 1$

ب یک عبارت توصیفی به‌عنوان مثال $\mu_i \in \{\cdots, low, medium, high, \cdots\}$ (توابع عضویت)

توابع فوق باید به‌نحوی بیان شوند که کل دامنه $(0, 1)$ را در برگیرند نسبت $\frac{\mu_i}{1-\mu_i}$ به‌عنوان بخت امکانی [1] برای فرد $i$ ام در نظر گرفته می‌شود و نسبت امکان موفقیت به امکان شکست بیان می‌کند.

به دلیل اهمیت مدل‌بندی لجستیک در داده‌های پزشکی، مبهم و نادقیق بودن یافته‌ها و حجم نمونه کم عمدتاً در این مطالعات با آن روبرو می‌شویم. هدف این فصل ارائه مدلی است که امکان ابتلا به بیماری را در مواردی که افراد مورد مطالعه مشکوک به بیماری هستند مدل‌بندی کند. این مدل رگرسیون لجستیک فازی است که در آن مقادیر متغیر پاسخ مبهم است و جای اعداد صفر و یک، مجموعه‌های فازی از بازه واحد هستند. در این مدل متغیرهای تبیینی، دقیق فرض می‌شوند. مدل مذکور قادر است امکان بیمار بودن فرد مشکوک به بیماری را در زمان حال و یا ابتلای او در آینده را بیان کند.

فرض کنید می‌خواهیم ابتلا به یک بیماری خاص را براساس مجموعه‌ای از عوامل خطر مدل‌بندی کنیم. اگر ابتلا به بیماری را متغیر پاسخ دودویی (بیمار/سالم) در نظر بگیریم، مدل‌بندی رابطه بین این متغیرها با استفاده از مدل رگرسیون لجستیک صورت می‌گیرد. برای برآورد پارامترهای مدل مذکور به نمونه‌ای به حجم $n$ نیاز داریم که شامل تعدادی افراد بیمار و سالم با مقادیر معلوم عوامل خطر است. معمولا پزشک در تشخیص وجود بیماری در فرد تردید دارد و امکان ابتلا به بیماری برای افراد مبهم است. در مدل‌های آماری این نمونه‌های مبهم از نمونه اصلی کنار گذاشته می‌شود و در مدل‌بندی شرکت داده نمی‌شوند، در واقع این

[1] Possibilistic odds

نمونه‌ها دور ریخته می‌شوند. بنابراین برای محاسبه احتمال $\pi = P(Y = ۱)$ از نظریه فازی کمک می‌گیریم. و همان‌طور که گفته شد به‌جای احتمال بیمار بودن، امکان ابتلا به بیماری براساس عوامل خطر را مدل‌بندی می‌کنیم. در مدل، امکان ابتلا را یک متغیر زبانی در نظر می‌گیریم. قبل از ارائه مدل، مفهوم متغیر زبانی [۲] را تعریف می‌کنیم.

**تعریف ۲.۱.۳.** [۱] یک متغیر زبانی توسط یک پنج تایی مرتب $(X, T(X), U, G, M)$ تعریف می‌شود که در آن $X$ نام متغیر است، $U$ مجموعه مرجع است، $T(X)$ مجموعه ترم (واژه)های مربوط به متغیر است و هر ترم یک مجموعه فازی است که توسط قاعده نحوی $G$ تولید می‌شود و سرانجام $M$ یک قاعده معنایی است که به هر ترم $T(X)$ معنای آن را مربوط می‌سازد، یعنی تابع عضویت آن ترم را مشخص می‌کند.

**مثال ۱.۱.۳.** متغیر سن را با مجموعه $U = [۰, ۲۰۰]$ در نظر بگیرید. مقادیر این متغیر اعدادی مانند ۳۵، ۷۰ و ۱۱۵ هستند. با این مقادیر، متغیر سن یک متغیر معمولی است. اما در زبان معمولی از مفاهیمی استفاده می‌کنیم که در واقع مفاهیم مبهمی از متغیر سن هستند، مانند خردسال، نوجوان، جوان، پیر و ... . در این مثال عنصر $X$ سن است، مجموعه مرجع $U = [۰, ۲۰۰]$ تعریف می‌شود، همچنین داریم:

$M$ مجموعه قواعدی است که به هر مقدار زبانی سن، معنای آن را نسبت می‌دهد. برای مثال ترم «جوان» را به‌صورت یک مجموعه فازی با تابع عضویت زیر می‌توان نوشت.

$$A(x) = \begin{cases} ۱ & ۰ \leq x < ۳۰ \\ ۱ - ۲\left(\frac{x - ۳۰}{۳۰}\right)^۲ & ۳۰ \leq x < ۴۵ \\ ۲\left(\frac{۶۰ - x}{۳۰}\right)^۲ & ۴۵ \leq x < ۶۰ \\ ۰ & ۶۰ \leq x \end{cases}$$

حال به کمک تعریف‌هایی که بیان شد به صورت‌بندی و برازش مدل و در نهایت برآورد پارامترهای مجهول مدل می‌پردازیم.

## ۱.۱.۳ برازش مدل رگرسیون لجستیک با رویکرد امکانی

در فصل قبل تاریخچه مختصری رگرسیون امکانی معرفی شد. در این قسمت به معرفی این روش می‌پردازیم.

در ابتدا از خبره می‌خواهیم با توجه به تجربه خود ترم‌های زبانی را برای هر موقعیت بیان کند، آنگاه ترم‌های زبانی را به‌صورت یک مجموعه فازی با تابع عضویتی بر بازه $(۰, ۱)$ تعریف می‌کنیم. آنگاه با استفاده از اصل گسترش، تبدیل یافته هر یک از ترم‌های زبانی را به‌عنوان

[۲] Linguistic Approach

ورودی‌های متغیر پاسخ در مدل پیشنهادی در نظر می‌گیریم. به‌عبارت دیگر $\tilde{\mu}_i$ یک متغیر زبانی است، که امکان ابتلا به بیماری را تعریف می‌کنیم. تبدیل یافته $\tilde{\mu}_i$ یعنی $\frac{\tilde{\mu}_i}{۱-\tilde{\mu}_i}$ را بخت امکانی می‌نامیم و لجیت آن، $\tilde{y}_i = \ln \frac{\tilde{\mu}_i}{۱-\tilde{\mu}_i}$ را به‌عنوان مشاهده متغیر پاسخ در نظر می‌گیریم. با استفاده از تبدیل لجیت $\tilde{y}_i$ به‌صورت خطی با پارامترهای مدل در ارتباط است و در نتیجه برای برآورد پارمترهای فازی مدل از رگرسیون خطی فازی می‌توان استفاده کرد. با توجه به تعریف $\tilde{y}_i$ و با استفاده از اصل گسترش، تابع عضویت $\tilde{y}_i$ عبارت است از

$$\tilde{y}_i(y) = \sup_{\forall x\in\mathbb{R}, x\neq ۱; \ln \frac{x}{۱-x}=y} \mu_i(x) \quad , i = ۱, ۲, \cdots, n \tag{۱.۳}$$

طبق معادله (۱.۳) $\tilde{y}_i$ یک عدد مثلثی متقارن و برآورد تبدیل لگاریتم بخت امکانی است. بنابراین داریم

$$\tilde{y}_i = (f_i^c(x), f_{is}^L(x), f_{is}^R(x))_T \tag{۲.۳}$$

که در آن

$$f_i^c(x) = a_۰^c + a_۱^c x_{i۱} + \cdot + a_{p-۱}^c x_{ip-۱}$$
$$f_{is}^L(x) = s_۰^L + s_۱^L x_{i۱} + \cdot + s_{p-۱}^L x_{ip-۱}$$
$$f_{is}^R(x) = s_۰^R + s_۱^R x_{i۱} + \cdot + s_{p-۱}^R x_{ip-۱}$$

به‌ترتیب مرکز، پهنای چپ و پهنای راست عدد فازی هستند. تابع عضویت برای برآورد فازی خروجی به‌صورت زیر است: [۴]

$$\tilde{Y}_i(\tilde{y}_i) = \begin{cases} ۱ - \dfrac{f_i^c(x) - y_i}{f_{is}^L(x)} & f_i^c(x) - f_{is}^L(x) \leq y_i \leq f_i^c(x) \\ ۱ - \dfrac{y_i - f_i^c(x)}{f_{is}^R(x)} & f_i^c(x) \leq y_i \leq f_i^c(x) + f_{is}^R(x). \end{cases} \tag{۳.۳}$$

اگر $s_i^L = s_i^R = s_i$ باشد، عدد فازی مثلثی متقارن خواهیم داشت. در این‌صورت برای $\hat{Y}_i$ شرط زیر را برای پهناهای چپ و راست عدد فازی خواهیم داشت.

$$f_i^L(x) = f_i^R(x) = f_i(x).$$

در این پایان‌نامه اعداد فازی مورد بررسی ما، اعداد فازی مثلثی نامتقارن هستند.

همان‌طور که می‌دانیم، لگاریتم بخت امکانی $\tilde{Y}_i$، شانس گرفتن یا داشتن ویژگی تعریف شده در عضو $i$ام داده‌ها می‌باشد. با توجه به اصل گسترش اگر $\tilde{M}$ یک عدد فازی با تابع عضویت $\tilde{Y}_i$ باشد و $f(x) = exp(x)$، در این صورت $f(\tilde{M}) = exp(\tilde{M})$ یک عدد فازی با تابع عضویت زیر است

$$exp(\tilde{M}(x)) = \begin{cases} \tilde{M}(\ln x) & x > ۰ \\ ۰ & o.w. \end{cases} \tag{۴.۳}$$

بنابراین، پس از برآورد ضرایب مدل، می‌توان تابع عضویت بخت امکانی را تعیین کرد.

(۵.۳)

$$exp(\tilde{Y}(x)) = \tilde{Y}_i(\ln(x)) = \begin{cases} ۱ - \dfrac{f_i^c(x) - \ln(x)}{f_{is}^L(x)} & f_i^c(x) - f_{is}^L(x) \le \ln(x) \le f_i^c(x) \\ ۱ - \dfrac{y_i - f_i^c(x)}{f_{is}^R(x)} & f_i^c(x) \le \ln(x) \le f_i^c(x) + f_{is}^R(x). \end{cases}$$

از این رو، برای یک مورد جدید فازی، مدل می‌تواند بخت امکانی خود را به عنوان یک عدد فازی با استفاده از برداری از مشاهدات ورودی ( مشاهدات تبیینی) دقیق پیش‌بینی کند.

## ۲.۱.۳ برآورد ضرایب فازی

در این بخش قصد داریم ضرایب $\tilde{b}_j, j = ۱, \cdots, n$ را با استفاده از رویکرد امکانی در مدل رگرسیون خطی فازی با متغیر پاسخ فازی، ضرایب فازی و متغیر تبیینی دقیق (غیر فازی) برآورد کنیم. ایده اساسی در این روش این است که مدل فازی به دست آمده را به حداقل برسانیم که این کار با به حداقل رساندن ضرایب فازی مقدور می‌شود. بنابراین فرض می‌شود که:

۱. برای مجموعه مشاهدات، $y_i$ دارای درجه عضویت به بزرگی $h$ در تابع خروجی برآورد فازی است. $\tilde{Y}_i(y_i) \ge h$ که ر آن

$$y_i = \ln(\frac{\mu_i}{۱ - \mu_i}), h \in (۰, ۱). \qquad (۶.۳)$$

۲. ضرایب فازی با به حداقل رساندن مدل فازی به دست می‌آیند. از آنجایی که فازی بودن اعداد فازی با پهنای آن افزایش می‌یابد،با به حداقل رساندن مجموع پهناهای خروجی فازی، مقدار فازی مدل حداقل می‌شود.

تعیین ضریب فازی منجر به یک مسئله برنامه‌ریزی خطی می‌شود که در آن تابع هدف مجموع پهناهای خروجی فازی است.

$$Z = n(s_۰^L + s_۰^R) + \sum_{j=۱}^{n}[(s_i^L + s_i^R)\sum_{i=۱}^{n} x_{ij}]. \qquad (۷.۳)$$

که در آن $x_{ij}$، $i$امین مقدار مشاهده شده برای $j$امین متغیر است.

از سوی دیگر، با توجه به معادله (۵.۳) هر ضابطه را با توجه به $\tilde{Y}_i(y_i) \ge h$ را می‌توان به صورت زیر نوشت.

$$۱ - \frac{f_i^c(x) - y_i}{f_{is}^L(x)} \ge h \Rightarrow (۱ - h)s_۰^L + (۱ - h)\sum_{j=۱}^{m} s_i^L x_{ij} - a_۰^c - \sum_{j=۱}^{m} a_j^c x_{ij} \ge -y_i$$

$$۱ - \frac{f_i^c(x) - y_i}{f_{is}^R(x)} \geq h \Rightarrow (۱-h)s_۰^R + (۱-h)\sum_{j=۱}^{m} s_i^R x_{ij} - a_۰^c - \sum_{j=۱}^{m} a_j^c x_{ij} \geq -y_i$$

با توجه به نامعادلات بالا، $۲m$ قید داریم. ازاین رو تابع هدف (۷.۳) را با استفاده از الگوریتم‌های برنامه‌نویسی خطی مانند سیمپلکس، به منظور برآورد مقدار مُد و پهناهای چپ و راست هر ضریب، به حداقل رساند.

## ۳.۱.۳ معیار نیکویی برازش

همانند سایر مدل‌سازی‌های آماری، مدل‌های مبتنی بر قوانین فازی، باید توسط برخی از روش‌ها (که بهترین مدل را برای داده‌ها برازش می‌کند) ارزیابی شوند. چندین روش خوب برای مدل‌های فازی ارائه شده است [۱۵] در اینجا دو روش برای ارزیابی مدل‌های رگرسیون لجستیک فازی پیشنهاد می‌کنیم.

### میانگین درجه عضویتMDM

**تعریف ۳.۱.۳.** مدل رگرسیون لجستیک فازی که براساس مشاهدات دقیق است را در نظر بگیرید. میانگین درجه عضویت (MDM) برای مقادیر مشاهده شده در تابع عضویت برآورد شده به صورت زیر تعریف می‌شود.

$$MDM = \frac{۱}{n}\sum_{i=۱}^{n} \tilde{Y}(y_i) = \frac{۱}{n}\sum_{i=۱}^{n} exp(\tilde{Y}_i(\frac{\mu_i}{۱-\mu_i})).$$

این یک معیار برای ارزیابی مدل است.

مقادیر بزرگ عضویت در مقادیر مشاهده شده تایید می‌کند که مدل ساخته شده از این داده‌ها، به خوبی از داده‌ها پشتیبانی می‌کند. بیشترین مقدار MDM برابر ۱ و کمترین مقدار آن ۰ است. بنابراین مقدار نزدیک به یک نشان می‌دهد که مدل مناسبی برازش داده شده است.

### میانگین مجموع توان‌های دوم خطای پیش بینی

روشی دیگر برای نیکویی برازش، اندازه‌گیری فاصله بین دو مقدار پاسخ مشاهده شده و پاسخ پیش‌بینی شده است. هر چه مقدار خروجی پیش‌بینی شده به مدل مشاهده شده نزدیکتر باشد، توان مدل برای پیش‌بینی وضعیت واقعی نمونه‌ها بالاتر می‌رود. با این حال، از آنجا که پاسخ مشاهده شده یک مقدار دقیق است، در حالی‌که مقدار پیش‌بینی یک عدد فازی می‌باشد، به همین علت لازم است که عدد فازی ابتدا با استفاده از روش‌های تبدیل فازی به دقیق، مقادیر پاسخ را به اعداد دقیق تبدیل کنیم[۱۳].

**تعریف ۴.۱.۳.** مدل رگرسیون لجستیک فازی با مشاهدات ورودی دقیق، $\frac{\mu_i}{1-\mu_i}$، و مقادیر برآورد شده فازی، $exp(\tilde{Y}_i)$، را در نظر بگیرید. میانگین توان دوم خطای پیشگویی، MSE، به صورت زیر به دست می‌آید.

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}[def(exp(\tilde{W}_i)) - \frac{\mu_i}{1-\mu_i}]$$

در معادله بالا، $def(exp(\tilde{W}_i))$ تبدیل شده فازی به دقیق [۳] از $exp(\tilde{Y}_i)$ است، که در تعریف زیر به آن پرداخته‌ایم.

**تعریف ۵.۱.۳.** [۴] فرض کنید که $\tilde{Y}$ یک عدد فازی در مجموعه $\mathbb{R}$ است، برای تبدیل عدد فازی $\tilde{Y}$ به عدد دقیق، به روش زیر عمل می‌کنیم. [۳۴]

$$def_{COG}(\tilde{Y}) = \frac{\int_x x\tilde{Y}dx}{\int_x \tilde{Y}dx} \qquad (۸.۳)$$

## ۲.۳ انتخاب متغیر در رگرسیون لجستیک فازی

معمولا در انتخاب مدل، دو معیار متقابل دخالت دارند.

۱. برای آن که معادله برای هدف‌های پیشگویی مفید باشد، ممکن است مایل باشیم مدل شامل حداکثر متغیرهای پیشگو باشد.

۲. به‌علت هزینه دستیابی به اطلاعات درباره تعداد زیاد متغیر پیشگو و ارائه آنها، ممکن است تمایل داشته باشیم معادله دارای حداقل تعداد متغیر پیشگو باشد.

شیوه یکتایی برای انتخاب مدل وجود ندارد. حتی معیارهای مختلف انتخاب مدل در یک مساله خاص هم لزوماً به یک مدل رگرسیونی یکتا و منحصر به‌فرد منجر نمی‌شود. حال برای ارزیابی صحیح رگرسیون لجستیک فازی برخی از معیارها را معرفی می‌کنیم.

همان‌طور که در فصل قبل بیان شد، مجموع مربعات خطا براساس فاصله بین مقادیر مشاهده شده و مقادیر پیشگویی شده محاسبه می‌شود. فاصله ژو (۴.۲)، یک فاصله رایج در تجزیه و تحلیل رگرسیون فازی است. در این فصل نیز برای انتخاب مدل مناسب از این فاصله استفاده می‌کنیم. میانگین توان‌های دوم خطا براساس متر ژو [۳۵] به صورت زیر تعریف می‌شود.

$$MPE = \frac{\sum_{i=1}^{n} d^2(\tilde{y}_i, \tilde{Y}_i)}{n} \qquad (۹.۳)$$

---

[۳] defuzzification

[۴] Center of Gravity Defuzzification Method

معیار اطلاعات آکائیک (AIC) در محیط فازی توسط آکائیک [۵] [۹]، تعریف شده است که معمولاً به‌عنوان یک معیار مناسب برای مقایسه مدل‌ها محسوب می‌شود. در تعریف زیر این معیار در محیط فازی و برای رگرسیون لجستیک فازی تعمیم داده شده است.

**تعریف ۱.۲.۳.** در محیط فازی، معیار ACI به‌صورت زیر تعریف می‌شود.

$$AIC = n\ln(SSE) - n\ln(n) - ۲(p+۱) \qquad (۱۰.۳)$$

که $n$ تعداد نمونه‌ها و $p$ تعداد متغیرهای مدل و $SSE = \sum_{i=۱}^{n} d^{۲}(\tilde{y}_i, \tilde{Y}_i)$ است.

**تعریف ۲.۲.۳.** در دامنه اعداد فازی، $C_p$ به‌صورت زیر تعریف می‌شود:

$$C_p = \frac{SSE_{model}}{SSE_{saturated.model}} + ۲(p+۳) - n \qquad (۱۱.۳)$$

که در آن n حجم نمونه، p تعداد ضرائب، $SSE_{model}$ مجموع مربعات خطای مدل مورد بررسی و $SSE_{saturated.model}$ مجموع مربعات مدل که شامل همه متغیرهای پیشگو می‌باشد.

## ۱.۲.۳ روش پیشرو برای انتخاب متغیر به شیوه سلمانی و همکاران

سلمانی و همکاران روش پیشرو برای انتخاب بهترین متغیر در رگرسیون لجستیک فازی را معرفی کردند. در روش پیشرو در حالت غیر فازی، متغیرهای پیشگو براساس کمترین میانگین مربعات خطا به مدل اضافه می‌شوند و تا زمانی که بهبود در مدل ایجاد شود روند اضافه شدن متغیرهای پیشگو به مدل ادامه دارد. مزیت این روش در این است که تعداد تکرار کمی دارد. آنها از معیار سطح اثر بخشی [۶] (LE) به‌عنوان یک معیار برای ارزیابی هر مدل در روش پیشرو استفاده کردند.

## ۲.۲.۳ معیار LE

اهمیت انتخاب سطح ورود و خروج به‌منظور قضاوت در مورد ارزش متغیرهای کمکی در روش انتخاب پیشرو مهم است. معیار سطح اثر بخشی برای تصمیم‌گیری در مورد وجود یا عدم وجود متغیرها در مدل معرفی شده‌است. با توجه به LE، متغیری که واردمدل می‌شود، می‌تواند حداقل درصد LE از مدل قبلی را بهبود بخشد. اگر سطح کارایی متغیر وارد شده به مدل کم باشد منجر به افزایش ورود متغیرها به مدل می‌شود. این پیچیدگی مدل را افزایش می‌دهد و آنالیز را دشوار می‌سازد، لذا پیش‌بینی با عدم اطمینان بیشتری انجام می‌شود. از سوی دیگر، افزایش تعداد متغیرها در مطالعات بالینی شامل صرف هزینه زیادی برای جمع‌آوری اطلاعات در مورد متغیرها می‌شود. بنابراین انتخاب این معیار بستگی زیادی به شرایط مطالعه دارد. سلمانی و همکاران [۲۷] ، یک الگوریتم پیشنهادی را برای انتخاب مدل مناسب معرفی کردند.

[۵] Akaike

[۶] Level of efficacy

### ۳.۲.۳ الگوریتم انتخاب متغیر به روش پیشرو

الگوریتم سلمانی و همکاران [۲۷] با این فرض آغاز می‌شود که هیچ متغیر تبیینی‌ای در مدل حضور نداشته باشد و فقط عرض از مبدا وجود دارد. اولین متغیر پیشگوی فازی که وارد مدل می‌شود، متغیری است که بیشترین تغییرات مدل را توصیف کرده باشد. در گام بعدی، شرایط هر متغیر برای وارد شدن به مدل، سطح کارایی حداقل LE است. گام‌های اساسی این شیوه عبارتند از: (شکل ۲.۳)

۱. مدل‌های تک متغیره برازش داده می‌شوند و از بین آنها بهترین مدل انتخاب می‌شود.

۲. متغیر بعدی وارد مدل می‌شود و MSE محاسبه می‌شوند. اگر درصد بهبود مدل بیشتر از LE درصد باشد، متغیر در مدل می‌ماند.

۳. این روند تا جایی که افزودن متغیر به مدل باعث تغییری در شاخص نشود، ادامه پیدا می‌کند.

### ۴.۲.۳ کاربرد رگرسیون لجستیک فازی در بیماری لوپوس

بیماری لوپوس ارتیماتوز سیستمیک [۷] که در علم پزشکی به اختصار SLE نامیده می‌شود، گونه‌ای بیماری خود ایمنی است که در آن سیستم دفاعی بدن، علیه ارگان‌ها و بافت‌های پیوندی خودی عمل کرده و به آنها آسیب می‌رساند. زنان در مقابل این بیماری ده برابر بیشتر از مردان آسیب پذیرند. این بیماری اغلب پوست و چندین اندام داخلی را درگیر می‌کند و با وجود اتوآنتی بادی در خون همراه است. سیر بالینی لوپوس شامل دوره‌های فعالیت و بهبود است. علت بیماری، نقص ایمنی پوست و ارگان‌های داخلی است که منشأ آن تاکنون شناخته شده نیست. نوع حاد آن به ندرت به مرگ منجر می‌شود و بیشتر در زنان و دردهه سوم زندگی بروز می‌کند. [۱۸] تشخیص بیماری براساس علائم آن صورت می‌گیرد. تشخیص در مراحل اولیه باعث تسریع در درمان و مانع از پیشرفت بیماری می‌شود.

داده‌ها متشکل از ۱۵ نفر زن مشکوک به بیماری لوپوس در بازه سنی ۴۰ – ۱۸ سال از درمانگاه‌های شیراز جمع‌آوری شده است. در این تحقق می‌خواهیم وصعیت افراد مشکوک (سالم/ بیمار) به بیماری لوپوس را براساس تعدادی عوامل خطر، مدل‌سازی کنیم، به گونه‌ای که مدل برازش داده شده قادر باشد بخت امکانی ابتلا به بیماری در هر فرد را برآورد کند. عوامل خطری که در این مطالعه استفاده شده است شامل تست‌های آزمایشگاهی ESR، Anti DNA، ANA، مواجه فرد با نور خورشید و سابقه خانوادگی فرد است. آزمایش‌هایی که نام برده شد، آزمایش‌های ویژه خونی هستند.

[۷] Systemic Lupus Erythematosus

با توجه به ترم‌های زبانی ((خیلی کم، کم، متوسط، زیاد، خیلی زیاد)) وضعیت عمومی هر یک از ۱۵ فرد توسط پزشک بیان شده است. در همین راستا توابع عضویت هر ترم زبانی به‌صورت زیر تعریف شدند.

$$VeryLow = \begin{cases} ۱ - \frac{۰/۰۲ - x}{۰/۰۱} & ۰/۰۱ \le x \le ۰/۰۲ \\ ۱ - \frac{x - ۰/۰۲}{۰/۱۶} & ۰/۰۲ \le x \le ۰/۱۸ \end{cases} \quad Low = \begin{cases} ۱ - \frac{۰/۲۵ - x}{۰/۱۵} & ۰/۱ \le x \le ۰/۲۵ \\ ۱ - \frac{x - ۰/۰۲}{۰/۱۵} & ۰/۲۵ \le x \le ۰/۴ \end{cases}$$

$$Medium = \begin{cases} ۱ - \frac{۰/۰۵ - x}{۰/۱۵} & ۰/۳۵ \le x \le ۰/۵ \\ ۱ - \frac{x - ۰/۵}{۰/۱۵} & ۰/۵ \le x \le ۰/۶۵ \end{cases} \quad High = \begin{cases} ۱ - \frac{۰/۷۵ - x}{۰/۱۵} & ۰/۶ \le x \le ۰/۷۵ \\ ۱ - \frac{x - ۰/۷۵}{۰/۱۵} & ۰/۷۵ \le x \le ۰/۹ \end{cases}$$

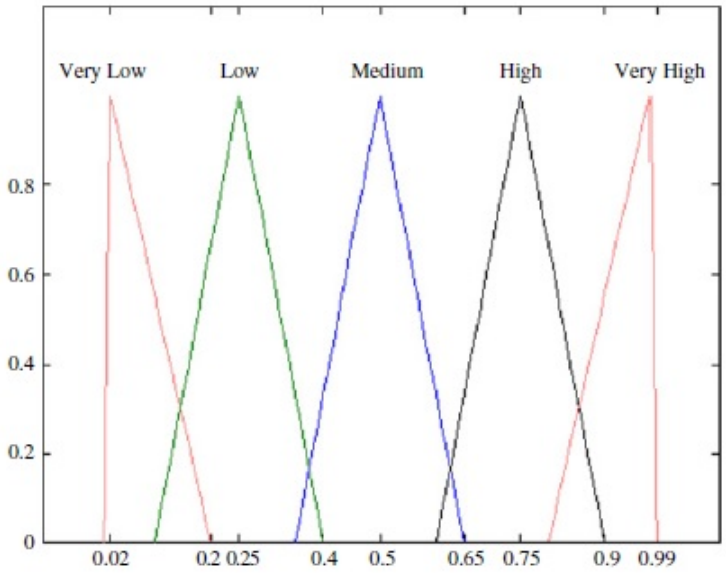
$$VeryHigh = \begin{cases} ۱ - \frac{۰/۹۸ - x}{۰/۱۸} & ۰/۸ \le x \le ۰/۹۸ \\ ۱ - \frac{x - ۰/۹۸}{۰/۰۱} & ۰/۹۸ \le x \le ۰/۹۹ \end{cases} \quad (۱۲.۳)$$

داده‌های این تحقیق در جدول ۱.۳ آمده است. مدل کلی با شرکت همه متغیرهای پیش‌پینی و پاسخ به‌صورت زیر است.

$$\hat{\tilde{Y}}_i = \tilde{b}_۰ + \tilde{b}_۱ ESRtest_i + \tilde{b}_۲ AntiDNAtest_i + \tilde{b}_۳ ANAtest_i + \tilde{b}_۴ SunExpossure_i + \tilde{b}_۵ FamilyHistory_i \quad (۱۳.۳)$$

همچین مقادیر مشاهده شده $\tilde{y}_i$ را با استفاده از توابع عضویت و تبدیل لجیت $\tilde{y}_i = \ln \frac{\mu_i}{۱-\mu_i}$ به دست آوردیم. برای این کار از نرم افزار R و از بسته FuzzyNumber کمک گرفتیم. کدها مربوط به R در پیوست ب.۱ و مقادیر به دست آمده در جدول ۲.۳ آوردیم.

با روش پیشرو و به کمک الگوریتمی که توضیح دادیم با معیار LE بهترین مدل را انتخاب می‌کنیم.




شکل ۱.۳: نمودار پیشنهادی $\mu_i$

جدول ۱.۳: موارد مشکوک به لوپوس و عوامل خطر مربوط به آن

| شماره فرد | ESR test | AntiDNA test | ANA test | مواجه با نور خورشید | سابقه خانوادگی | امکان $\mu_i$ |
|---|---|---|---|---|---|---|
| ۱ | ۱ | ۱۰۵ | ۱۱۲ | ۱ | ۱ | high |
| ۲ | ۰ | ۲۳ | ۸۰ | ۱ | ۰ | medium |
| ۳ | ۰ | ۱۵ | ۱۱۵ | ۱ | ۰ | high |
| ۴ | ۱ | ۱۰۷ | ۱۰۵ | ۱ | ۰ | high |
| ۵ | ۱ | ۱۵۰ | ۸۹ | ۰ | ۰ | medium |
| ۶ | ۱ | ۱۰ | ۱۶۰ | ۱ | ۱ | veryhigh |
| ۷ | ۰ | ۲۳ | ۱۰۰ | ۱ | ۰ | medium |
| ۸ | ۱ | ۸۵ | ۱۰۰ | ۰ | ۰ | high |
| ۹ | ۰ | ۸۳ | ۴۸ | ۱ | ۰ | low |
| ۱۰ | ۱ | ۱۹ | ۱۵ | ۰ | ۱ | low very |
| ۱۱ | ۰ | ۹۱ | ۵۰ | ۰ | ۰ | low |
| ۱۲ | ۱ | ۲۰۰ | ۵۹ | ۱ | ۰ | medium |
| ۱۳ | ۱ | ۲۰ | ۸۳ | ۱ | ۰ | low |
| ۱۴ | ۰ | ۲۰۰ | ۱۵ | ۰ | ۰ | low |
| ۱۵ | ۱ | ۱۵ | ۸۵ | ۰ | ۱ | medium |

جدول ۲.۳: مقادیر مشاهده شده $\tilde{y}_i$

| ردیف | مرکز | پهنای چپ | پهنای راست |
|---|---|---|---|
| ۱ | ۱.۰۶۳۷۲۱ | ۰.۶۵۱۷۵۲۱ | ۱.۰۳۵۹۳۴۵ |
| ۲ | ۰.۰۰۰۰۰۰ | ۰.۶۰۹۳۴۱۵ | ۰.۶۰۹۳۴۱۵ |
| ۳ | ۱.۰۶۳۷۲۱ | ۰.۶۵۱۷۵۲۱ | ۱.۰۳۵۹۳۴۵ |
| ۴ | ۱.۰۶۳۷۲۱ | ۰.۶۵۱۷۵۲۱ | ۱.۰۳۵۹۳۴۵ |
| ۵ | ۰.۰۰۰۰۰۰ | ۰.۶۰۹۳۴۱۵ | ۰.۶۰۹۳۴۱۵ |
| ۶ | ۳.۳۱۷۴۰۳ | ۲.۱۶۳۹۷۰۲ | ۱.۲۳۲۵۵۵۲ |
| ۷ | ۰.۰۰۰۰۰۰ | ۰.۶۰۹۳۴۱۵ | ۰.۶۰۹۳۴۱۵ |
| ۸ | ۱.۰۶۳۷۲۱ | ۰.۶۵۱۷۵۲۱ | ۱.۰۳۵۹۳۴۵ |
| ۹ | −۱٫۰۶۳۷۲۱ | ۱.۰۳۵۹۳۴۵ | ۰.۶۵۱۷۵۲۱ |
| ۱۰ | −۳٫۸۵۷۵۶۶ | ۰.۶۹۲۳۹۱۷ | ۲.۵۶۱۲۲۹۳ |
| ۱۱ | −۱٫۰۶۳۷۲۱ | ۱.۰۳۵۹۳۴۵ | ۰.۶۵۱۷۵۲۱ |
| ۱۲ | ۰.۰۰۰۰۰۰ | ۰.۶۰۹۳۴۱۵ | ۰.۶۰۹۳۴۱۵ |
| ۱۳ | −۱٫۰۶۳۷۲۱ | ۱.۰۳۵۹۳۴۵ | ۰.۶۵۱۷۵۲۱ |
| ۱۴ | −۱٫۰۶۳۷۲۱ | ۱.۰۳۵۹۳۴۵ | ۰.۶۵۱۷۵۲۱ |
| ۱۵ | ۰.۰۰۰۰۰۰ | ۰.۶۰۹۳۴۱۵ | ۰.۶۰۹۳۴۱۵ |

جدول ۳.۳: مرحله اول انتخاب بهترین مدل

| شماره | مدل | $MSE$ |
|---|---|---|
| ۱ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ESRtest_i$ | ۹۵.۴۹۰۳ |
| ۲ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 AntiDNAtest_i$ | ۹۸.۷۷۴۹ |
| ۳ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i$ | ۱۸.۸۱۵۶ |
| ۴ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 SunExpossure_i$ | ۸۲.۲۱۹ |
| ۵ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 FamilyHistory_i$ | ۹۸.۲۱۹۳ |

### انتخاب بهترین مدل

**مرحله۱.** طبق الگوریتمی که گفته شد، برای تمامی مدل‌ها با یک متغیر، $MSE$ آنها را حساب می‌کنیم. همه مقادیر محاسبه شده را در جدول ۳.۳ آورده‌ایم. بهترین مدل در این مرحله

جدول ۴.۳: مرحله دوم انتخاب بهترین مدل

| شماره | مدل | $MSE$ |
|---|---|---|
| ۱ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 ESRtest_i$ | ۱۸.۳۰۵۵ |
| ۲ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 AntiDNAtest_i$ | ۹.۵۹۵۹ |
| ۳ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 SunExpossure_i$ | ۱۸.۷۶۷۱ |
| ۴ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 FamilyHistory_i$ | ۱۷.۵۷۰۰ |

$\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i$ است که کمترین $MSE$ با مقدار ۱۸/۸۱۵۶ را دارد.

**مرحله ۲.** در مرحله بعد با حفظ متغیر $ANAtest$ در مدل، متغیرهای دیگر را یک به یک وارد مدل کردیم. با توجه به مقدار $MSE$، بهترین مدل $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 AntiDNAtest_i$ است. اطلاعات مربوط به این مرحله در جدول ۴.۳ آمده است.

**مرحله۳.** این کار را تا جایی ادامه دادیم که با اضافه شدن متغیر جدید در مدل، هیچ بهبود در آن ایجاد نشود. اطلاعات مربوط به این مراحل در جداول ۵.۳ و ۶.۳ آمده است. در نهایت بهترین مدلی که با معیار $LE = 0/7$ انتخاب می‌کنیم مدل زیر است.

$$\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 AntiDNAtest_i + \tilde{b}_3 ESRtest_i + \tilde{b}_4 FamilyHistory_i \quad (۱۴.۳)$$

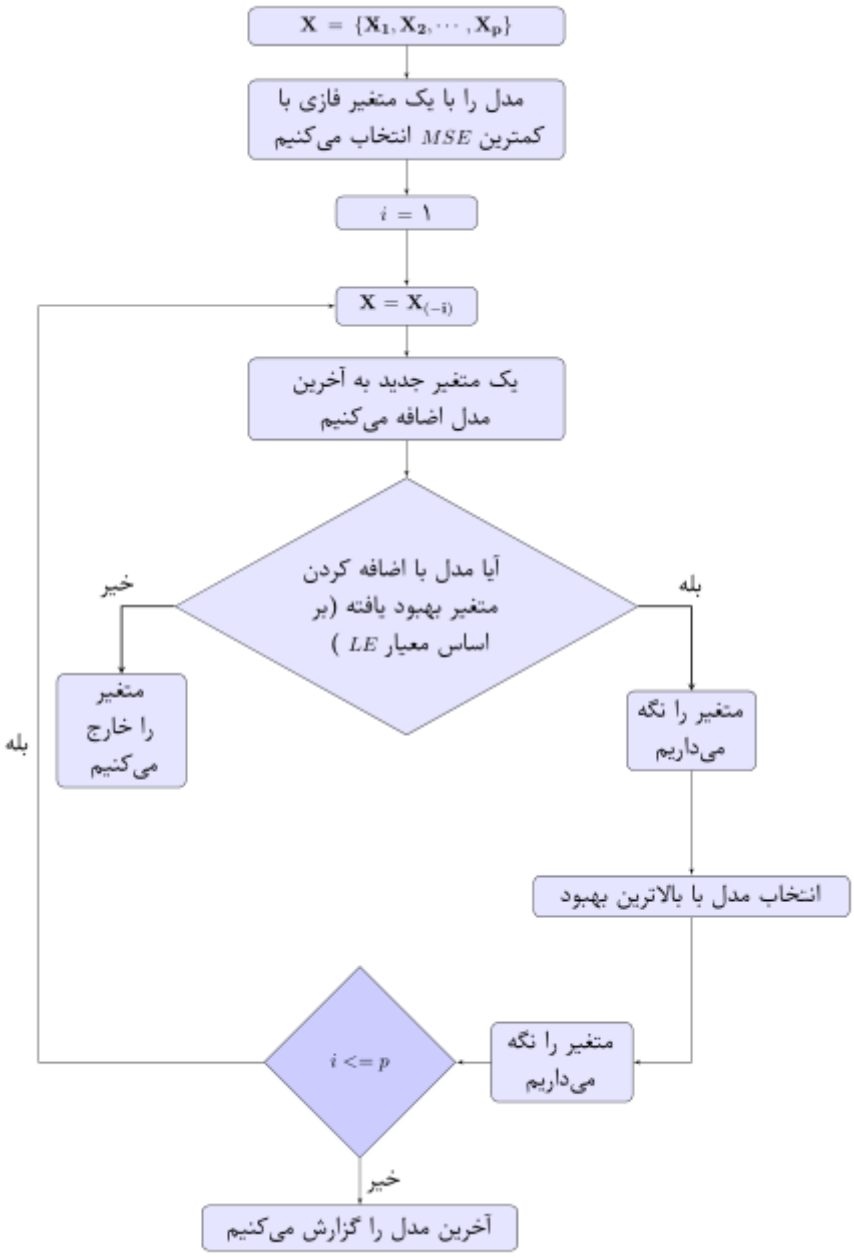
مدل (۱۴.۳) نشان می‌دهد که تست های آزمایشگاهی $ANA$، $AntiDNA$ و $ESR$ و سابقه خانوادگی عوامل تاثیر گذار در امکان ابتلای افراد به بیماری لوپوس را دارند. بنابراین $\tilde{Y}_i$ لگاریتم بخت امکانی بیماری برای فرد $i$ام است. با استفاده از اصل گسترش و با استفاده از تابع آنتی لگاریتم $f(\tilde{Y}_i) = e^{(\tilde{Y}_i)}$ امکان ابتلای هر فرد به بیماری را می‌توان محاسبه کرد.

جدول ۵.۳: مرحله سوم انتخاب بهترین مدل

| شماره | مدل | $MSE$ |
|---|---|---|
| ۱ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 AntiDNAtest_i + \tilde{b}_3 ESRtest_i$ | ۷.۹۹۶۳ |
| ۲ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 AntiDNAtest_i + \tilde{b}_3 SunExpossure_i$ | ۹.۵۸۱۹ |
| ۳ | $\tilde{Y}_i = \tilde{b}_0 + \tilde{b}_1 ANAtest_i + \tilde{b}_2 AntiDNAtest_i + \tilde{b}_3 FamilyHistory_i$ | ۹.۳۰۴۹ |

جدول ۶.۳: مرحله چهارم انتخاب بهترین مدل

| شماره | مدل | $MSE$ |
|---|---|---|
| ۱ | $\tilde{Y}_i = \tilde{b}_۰ + \tilde{b}_۱ ANAtest_i + \tilde{b}_۲ AntiDNAtest_i + \tilde{b}_۳ ESRtest_i + \tilde{b}_۴ SunExpossure_i$ | ۹.۵۸۱۹ |
| ۲ | $\tilde{Y}_i = \tilde{b}_۰ + \tilde{b}_۱ ANAtest_i + \tilde{b}_۲ AntiDNAtest_i + \tilde{b}_۳ ESRtest_i + \tilde{b}_۴ FamilyHistory_i$ | ۹.۳۰۴۹ |




شکل ۲.۳: الگوریتم انتخاب متغیر سلمانی در مدل رگرسیون لجستیک فازی

# فصل ۴

# رگرسیون لجستیک فازی با رویکرد کمترین توان دوم

در این فصل برای برآورد پارامترهای رگرسیون لجستیک با رویکرد کمترین توان دوم، از دو متر، یکی متر ژو [۳۵] با استفاده از مقاله پوراحمد و همکاران [۳۲] و متر دورسو [۱۱] استفاده می‌کنیم. سپس به کمک مثال عددی و با استفاده از نرم افزار R و Mathematica مطلب را کامل‌تر توضیح می‌دهیم.

## ۱.۴ روش مجموع توان دوم برای داده‌های فازی

رویکرد کمترین توان دوم فازی که اولین بار توسط دیاموند [۱۲] و کلمینس [۱۰] به طور هم‌زمان پیشنهاد شد، این روش در واقع یک تعمیم فازی از روش کمترین توان دوم با داده‌های دقیق است و مستلزم یک مفهوم فاصله بین مقادیر فازی مشاهده شده و مقادیر فازی برآورد شده است. همان‌طور که بیان شد در روش مجموع توان دوم، فاصله بین مقادیر مشاهده شده و مقادیر برآورد شده فازی را به حداقل می‌رسانیم که مزیت این روش است، زیرا باقیمانده‌ها برای سنجش دقت مدل مناسب هستند.

روش کمترین توان دوم خطی برای $m$ متغیر دقیق $(x_i, y_j)$، $1 \leq m \leq i$ برای پیدا کردن

مقادیر مجهول $a, b \in \mathbb{R}$ با حداقل کردن مجموع زیر است.

$$r(a,b) = \sum_i (a + bx_i + y_i)^۲. \tag{۱.۴}$$

در مواقعی که داده‌ها فازی هستند ما برای اعداد فازی مانند $\tilde{A}_۱$ و $\tilde{A}_۲$ به دنبال آن هستیم تا فاصله بین این دو عدد را حداقل کنیم.

### ۱.۱.۴ برآورد پارامترها با متر ژو

برای تعریف بهترین مدل بهینه و برای برآورد پارامترهای مدل از روش کمترین توان دوم خطا استفاده می‌کنیم. در این روش، مجموع توان دوم خطای بین $\tilde{y}_i$ و $\tilde{Y}_i$ را باید حداقل کرد. یعنی مجموع توان دوم خطا بین مقدار مشاهده شده و مقادیر برآورد شده را حداقل می‌کنیم. برای این منظور $SSE$ به‌صورت زیر تعریف می‌شود

$$SSE = \sum_{i=۱}^{n} d^۲(\tilde{Y}_i, \tilde{y}_i) \tag{۲.۴}$$

در این تعریف $d^۲(\tilde{Y}_i, \tilde{y}_i)$ فاصله بین دو مجموعه فازی است. یک تابع فاصله مناسب، تابع وزنی است که توسط ژو [۳۵] در رابطه (۴.۲) معرفی شده است.

همان‌طور که گفته شد مدل پیشنهادی رگرسیون لجستیک فازی

$$\tilde{Y}_i = \tilde{A}_۰ + \tilde{A}_۱ x_{i۱} + \cdots + \tilde{A}_{p-۱} x_{ip-۱} \tag{۳.۴}$$

می‌باشد، که در آن متغیر پاسخ فازی و متغیر تبیینی دقیق است و همچنین ضرایب مدل فازی هستند. گفته شد که برای برآورد ضرائب باید مجموع توان دوم خطاها (۲.۴) حداقل شود.

حال فرض کنید $\tilde{A}_j = (a_j, s_j^L, s_j^R)_T, j = ۱, \cdots, p-۱$ ضرایب مدل (۳.۴) باشد، در نتیجه پاسخ برآورد شده، مقدار عدد فازی مثلثی نامتقارن $\tilde{Y}_i = (f_i(a), f_i^L(s), f_i^R(s))_T$ است که در آن مرکز، پهنای چپ و پهنای راست به ترتیب

$$\begin{aligned} f_i^a(x) &= a_۰^c + a_۱^c x_{i۱} + \cdots + a_{p-۱}^c x_{ip-۱} \\ f_i^L(s) &= s_۰^L + s_۱^L x_{i۱} + \cdots + s_{p-۱}^L x_{ip-۱} \\ f_s^R(s) &= s_۰^R + s_۱^R x_{i۱} + \cdots + s_{p-۱}^R x_{ip-۱} \end{aligned} \tag{۴.۴}$$

می‌باشند.

هدف ما به دست آوردن فاصله بین $(\tilde{Y}_i)_\alpha$ و $(\tilde{y}_i)_\alpha$، یعنی فاصله بین مقادیر برآورد شده و مقادیر مشاهده شده و حداقل سازی بین این دو فاصله است. یعنی در نهایت مقدار زیر را به دست آورده تا بتوان پارامترهای مجهول مدل را برآورد کرد. در رابطه (۴.۲) :

$$d(\tilde{y}_i, \tilde{Y}_i) = [\int_۰^۱ f(\alpha) d^۲((\tilde{Y}_i)_\alpha, (\tilde{y}_i)_\alpha) d\alpha]^{\frac{۱}{۲}} \tag{۵.۴}$$

برای حالتی که ضرایب اعداد فازی مثلثی نامتقارن هستند، مقدار $\alpha-$برش $(\tilde{Y}_i)_\alpha$ به صورت

$$(\tilde{Y}_i)_\alpha = [f_i(\mathbf{a}) - (۱-\alpha)f_i^L(\mathbf{s}), f_i(\mathbf{a}) + (۱-\alpha)f_i^R(\mathbf{s})] \qquad (۶.۴)$$

به دست می‌آید. همچنین $\alpha-$ برش $(\tilde{y}_i)_\alpha$ براساس $(\tilde{\mu}_i)_\alpha = [b_۱, b_۲]$ به صورت زیر:

$$(\tilde{y}_i)_\alpha = [\ln \frac{b_۱}{۱-b_۱}, \ln \frac{b_۲}{۱-b_۲}]$$

می‌باشد.مقدار $b_۱$ و $b_۲$ از $\alpha-$ برش توابع عصویت براساس توابع عضویت بیان شده در (۱۳.۳) به صورت زیر به دست می‌آید:

$$(\tilde{\mu}_i)_\alpha = [f_i^L(\mathbf{s}) + (\alpha-۱)f_i(\mathbf{a}), f_i^R(\mathbf{s}) + (۱-\alpha)f_i(\mathbf{a})]$$

بنابراین فاصله بین $(\tilde{Y}_i)_\alpha$ و $(\tilde{y}_i)_\alpha$ بر اساس رابطه (۵.۲) به صورت

(۷.۴)

$$d^۲((\tilde{y}_i)_\alpha, (\tilde{Y}_i)_\alpha) = [\ln \frac{b_۱}{۱-b_۱} - f_i(\mathbf{a}) + (۱-\alpha)f_i^L(\mathbf{s})]^۲ + [\ln \frac{b_۲}{۱-b_۲} - f_i(\mathbf{a}) - (۱-\alpha)f_i^R(\mathbf{s})]^۲$$

خواهد بود. بنابراین داریم:

(۸.۴)

$$SSE = \sum_{i=۱}^{n} \int_۰^۱ f(\alpha)[[\ln \frac{b_۱}{۱-b_۱} - f_i(\mathbf{a}) + (۱-\alpha)f_i^L(\mathbf{s})]^۲ + [\ln \frac{b_۲}{۱-b_۲} - f_i(\mathbf{a}) - (۱-\alpha)f_i^R(\mathbf{s})]^۲]d\alpha$$

این تابع از طریق $f_i(\mathbf{a})$، $f_i^L(\mathbf{s})$ و $f_i^R(\mathbf{s})$ به ضرایب مدل وابسته است. با حداقل کردن (۸.۴) مقادیر مجهول $a_j$، $s_j^L$ و $s_j^R$ با شرط مثبت بودن پهناهای چپ و پهناهای راست، یعنی $s_j^L > ۰$ و $s_j^R > ۰$ به دست می‌آید.

برای حالتی که $s_j = s_j^L = s_j^R$ باشد، پوراحمد و همکاران [۳۲]

برای به دست آوردن پارامترها از $SSE$ نسبت به $a_j$ و $s_j$ مشتق گرفته شده است.

(۹.۴)

$$SSE = \sum_{i=۱}^{n} \int_۰^۱ f(\alpha)\Big[[\ln \frac{b_۱}{۱-b_۱} - f_i(\mathbf{a}) + (۱-\alpha)f_i(\mathbf{s})]^۲ + [\ln \frac{b_۲}{۱-b_۲} - f_i(\mathbf{a}) - (۱-\alpha)f_i(\mathbf{s})\Big]d\alpha.$$

با در نظر گرفتن $f(\alpha) = \alpha$ و

$$\frac{\partial SSE}{\partial a_j} = ۰ \qquad \frac{\partial SSE}{\partial s_j} = ۰$$

داریم:

$$\sum_{i=۱}^{n} \left( \int_۰^۱ ۲\alpha x_{ij} \left[ ۲f_i(\mathbf{a}) - \ln \frac{b_۱}{۱-b_۱} - \ln \frac{b_۲}{۱-b_۲} \right] d\alpha \right) = ۰ \quad j = ۰, ۱, \cdots, p-۱$$

$$\sum_{i=۱}^{n} \left( \int_۰^۱ ۲\alpha(۱-\alpha) x_{ij} \left[ ۲(۱-\alpha)f_i(\mathbf{a}) - \ln \frac{b_۱}{۱-b_۱} - \ln \frac{b_۲}{۱-b_۲} \right] d\alpha \right) = ۰ \quad j = ۰, ۱, \cdots, p-۱$$

مقدار معادلات فوق به تعریف امکان موفقیت $(\mu_i)$ برای هر فرد توسط خبره بستگی دارد. با جایگذاری این مقادیر و محاسبه انتگرال مربوط، معادلات زیر حاصل می‌شوند.

$$a_۰\sum_{i=۱}^{n}x_{i۰}x_{ij}+a_۱\sum_{i=۱}^{n}x_{i۱}x_{ij}+\cdots+a_{p-۱}\sum_{i=۱}^{n}x_{ip-۱}x_{ij}=\sum_{i=۱}^{n}z_{ij}x_{ij}$$
$$s_۰\sum_{i=۱}^{n}x_{i۰}x_{ij}+s_۱\sum_{i=۱}^{n}x_{i۱}x_{ij}+\cdots+s_{p-۱}\sum_{i=۱}^{n}x_{in}x_{ij}=\sum_{i=۱}^{n}k_{ij}x_{ij}$$

اگر $x_{i۰}=۱$ و $z_{ij}$ و $k_{ij}$ نتایج محاسبه انتگرال برای هر مورد باشد، بنابراین می‌توانیم معادلات بالا را به صورت ساده‌تر نوشت.

$$A\mathbf{a}=Z \quad , \quad A\mathbf{s}=K$$

که در آن

$$X=\begin{bmatrix}۱ & x_{۱۱} & \cdots & x_{۱p-۱}\\ ۱ & x_{۲۱} & \cdots & x_{۲p-۱}\\ \vdots & \vdots & & \vdots\\ ۱ & x_{n۱} & \cdots & x_{np-۱}\end{bmatrix}, \qquad A=X^\top X$$

$$\mathbf{a}=(a_۰,a_۱,\cdots,a_{p-۱})^\top \quad , \quad \mathbf{s}=(s_۰,s_۱,\cdots,s_{p-۱})^\top$$

$$Z=\left(\sum_{i=۱}^{n}z_ix_{i۰},\sum_{i=۱}^{n}z_ix_{i۱},\cdots,\sum_{i=۱}^{n}z_ix_{ip-۱}\right)^\top,$$
$$K=\left(\sum_{i=۱}^{n}k_ix_{i۰},\sum_{i=۱}^{n}k_ix_{i۱},\cdots,\sum_{i=۱}^{m}k_ix_{ip-۱}\right)^\top.$$

هستند. در نهایت مقادیر مجهول به صورت $\mathbf{a}=A^{-۱}Z$ و $\mathbf{s}=A^{-۱}K$ به دست می‌آید.

همان طور که بیان شده، در این پایان‌نامه، داده‌ها به صورت اعداد فازی مثلثی نامتقارن مفروضند. برای برآورد پارامترهای مجهول بایستی (۸.۴) را حداقل کنیم، یعنی نسبت به $a_j$، $s_j^L$ و $s_j^R$ مشتق گرفته و مساوی صفر قرار دهیم. در این صورت با عملیات ریاضی دشواری مواجه می‌شویم. برای حداقل کردن مجموع مربعات از (۱۰.۲) استفاده می‌نماییم. در ادامه با ذکر مثال، به شرح آن می‌پردازیم.

### ۲.۱.۴ مثال کاربردی

داده‌های بیماری لوپوس در بخش ۴.۲.۳ را در نظر بگیرید. برای برآورد پارامترهای مجهول مدل (۳.۴) در حالت نامتقارن بودن اعداد فازی مثلثی، باید مجموع مربعات (۸.۴) را حداقل

کنیم. برای این منظور از نرم افزار Mathematica استفاده کردیم، که کدهای آن در پیوست ب.۲ آورده شده است. مدل برازش شده به صورت زیر به دست می‌آید.

$$\begin{aligned}\hat{\tilde{Y}}_i = \ln\left(\frac{\tilde{\mu}}{۱-\tilde{\mu}}\right) &= (۰/۴۷۴, ۰/۹۹۶, ۱/۱۱۳) + (۰/۳۴۸, ۱/۰۹۹, ۱/۲۰۷)ESRtest_i \\ &+ (۰/۵۵۹, ۱/۳۲۴, ۰/۴۴۵)AntiDNAtest_i + (۰/۵۱۷, ۱/۴۸۸, ۱/۵۴۴)ANAtest_i \\ &+ (-۰/۵۴۵, ۰/۸۰۷, ۰/۶۶۶)SunExpossure_i + (۰/۳۲۴, ۱/۱۷۸, ۰/۴۶۶)FamilyHistory_i \end{aligned} \quad (۱۰.۴)$$

با استفاده از مدل (۱۰.۴) می‌توان امکان وجود بیماری در یک فرد را پیش بینی کرد. برای مثال با توجه به داده‌های جدول ۱.۳ برای فرد هشتم، از رابطه (۱۰.۴) داریم:

$$\begin{aligned}\hat{\tilde{Y}}_۸ = \ln\left(\frac{\tilde{\mu}_۸}{۱-\tilde{\mu}_۸}\right) &= (۰/۴۷۴, ۰/۹۹۶, ۱/۱۱۳) + (۰/۳۴۸, ۱/۰۹۹, ۱/۲۰۷) \times ۱ \\ &+ (۰/۵۵۹, ۱/۳۲۴, ۰/۴۴۵) \times ۸۵ + (۰/۵۱۷, ۱/۴۸۸, ۱/۵۴۴) \times ۱۰۰ \\ &+ (-۰/۵۴۵, ۰/۸۰۷, ۰/۶۶۶) \times ۰ + (۰/۳۲۴, ۱/۱۷۸, ۰/۴۶۶) \times ۰ \\ &= (۱۰۰/۰۳۷, ۲۶۳/۴۳۵, ۱۹۴/۵۴۵)\end{aligned}$$

بنابراین، لگاریتم بخت امکانی بیماری برای فرد هشتم، ۱۰۰/۰۳۷ برآورد می‌شود و با استفاده از اصل گسترش و با تبدیل $f(x) = e^x$، امکان بیماری برای فرد به صورت زیر محاسبه می‌شود.

$$\begin{aligned}\left(\frac{\tilde{\mu}_۸}{۱-\tilde{\mu}_۸}\right)(x) = exp(\tilde{Y}_۸(x)) = \tilde{Y}_۸(\ln x), \quad x > ۰ \\ = \begin{cases} ۱ - \dfrac{۱۰۰/۰۳۷ - \ln x}{۲۶۳/۴۳۵} & ۱۶۳/۳۹۸ \le \ln x \le ۱۰۰/۰۳۷ \\ ۱ - \dfrac{\ln x - ۱۰۰/۰۳۷}{۱۹۴/۵۴۵} & ۱۰۰/۰۳۷ \le \ln x \le ۲۹۴/۵۸۲ \end{cases}\end{aligned}$$

از مدل (۱۰.۴) برای پیش بینی امکان وجود بیماری در افراد جدید نیز می‌توان استفاده کرد.

مدل برازش شده در پور احمد و همکاران [۳۲] با فرض متقارن بودن اعداد فازی مثلثی ضرایب و پاسخ برآورد شده به صورت

$$\begin{aligned}\hat{\tilde{Y}}_i =& (-۳/۸۵۹۱, ۱/۱۶۱۷) + (-۰/۶۰۸۳, ۰/۱۴۵۱)ESRtest_i + ۰/۰۰۹۱AntiDNAtest_i \\ &+ ۰/۰۴۳۱ANAtest_i + (-۰/۱۳۰۹, ۰/۰۳۶۶)SunExpossure_i \\ &+ (۰/۴۲۴۸, ۰/۳۸۳۲)FamilyHistory_i \end{aligned} \quad (۱۱.۴)$$

است. به طور مثال برای فرد سوم، از رابطه (۱۰.۴) داریم:

$$\begin{aligned}\hat{\tilde{Y}}_{۳} &= \ln\left(\frac{\tilde{\mu}_{۳}}{۱-\tilde{\mu}_{۳}}\right) = (-۳/۸۵۹۱, ۱/۱۶۱۷) + (-۰/۶۰۸۳, ۰/۱۴۵۱) \times ۰ \\ &+ ۰/۰۰۹۱ \times ۱۵ + ۰/۰۴۳۱ \times ۱۱۵ \\ &+ (-۰/۱۳۰۹, ۰/۰۳۶۶) \times ۱ + (۰/۴۲۴۸, ۰/۳۸۳۲) \times ۰ \\ &= (۱/۱۰۳۰, ۱/۱۹۱۳) = (۱/۱۰, ۱/۲۰)\end{aligned}$$

بنابراین، لگاریتم بخت امکانی بیماری برای فرد سوم، با استفاده از اصل گسترش، امکان بیماری برای فرد به صورت زیر محاسبه می‌شود.

$$\begin{aligned}\left(\frac{\tilde{\mu}_{۳}}{۱-\tilde{\mu}_{۳}}\right)(x) &= exp(\tilde{Y}_{۳}(x)) = \tilde{Y}_{۳}(\ln x), \quad x > ۰ \\ &= \begin{cases} ۱ - \dfrac{۱/۱۰ - \ln x}{۱/۲۰} & -۰/۲۴ \le \ln x \le ۱/۱۰ \\ & (۰/۷۹ \le x \le ۳/۰۰) \\ ۱ - \dfrac{\ln x - ۱/۱۰}{۱/۲۰} & ۱/۱۰ \le \ln x \le ۲/۴۵ \\ & (۳/۰۰ \le x \le ۱۱/۵۹) \end{cases}\end{aligned}$$

حال از مدل (۱۱.۴) برای پیش بینی امکان وجود بیماری در فرد جدید استفاده می‌کنیم. برای مثال فرض می‌کنیم برای یک فرد جدید مقادیر متغیرهای تبیینی، به صورت زیر مشاهده شده باشند.

$$\begin{aligned}&ESRtest_{new} = ۰, AntiDNAtest_{new} = ۲۰, ANAtest_{new} = ۱۰۰ \\ &SunExpossure_{new} = ۱, FamilyHistory_{new} = ۱\end{aligned}$$

در این صورت با استفاده از مدل (۱۱.۴) داریم:

$$\begin{aligned}\hat{\tilde{Y}}_{new} &= \ln\left(\frac{\tilde{\mu}_{new}}{۱-\tilde{\mu}_{new}}\right) = (-۳/۸۵۹۱, ۱/۱۶۱۷) + (-۰/۶۰۸۳, ۰/۱۴۵۱) \times ۰ \\ &+ ۰/۰۰۹۱ \times ۲۰ + ۰/۰۴۳۱ \times ۱۰۰ \\ &+ (-۰/۱۳۰۹, ۰/۰۳۶۶) \times ۱ + (۰/۴۲۴۸, ۰/۳۸۳۲) \times ۱ \\ &= (۰/۹۲۶۸, ۱/۵۸۱۵)\end{aligned}$$

بنابراین، لگاریتم بخت امکانی بیماری برای این فرد، حدوداً ۰/۹۳ برآورد می‌شود. تابع عضویت

برای فرد جدید با توجه به اصل گسترش به صورت زیر است:

$$\left(\frac{\tilde{\mu}_{new}}{۱-\tilde{\mu}_{new}}\right)(x) = exp(\tilde{Y}_{new}(x)) = \tilde{Y}_{new}(\ln x), \quad x > ۰$$

$$= \begin{cases} ۱ - \dfrac{۰/۹۲۶۸ - \ln x}{۱/۵۸۱۵} & -۰/۶۵۴۷ \leq \ln x \leq ۰/۹۲۶۸ \\ & ۰/۵۱۹۶ \leq x \leq ۲/۵۲۶۴ \\ ۱ - \dfrac{\ln x - ۰/۹۲۶۸}{۱/۵۸۱۵} & ۰/۹۲۶۸ \leq \ln x \leq ۲/۵۰۸۳ \\ & ۲/۵۲۶۴ \leq x \leq ۱۲/۲۸۴۰ \end{cases}$$

در ادامه این فصل به برآورد پارامترهای رگرسیون لجستیک در محیط فازی به روشی دیگر می‌پردازیم. روش ارائه شده در فصل قسمت برای برآورد پارمترهای رگرسیون لجستیک فازی با استفاده از فاصله ژو [۳۵] بود، در این فصل می‌خواهیم پارامترهای مدل لجستیک فازی را با استفاده از فاصله تعریف شده به شیوه کپی و دورسو [۱۱] برآورد کنیم.

## ۲.۴ روش پیشنهادی برای برآورد پارامترها

برای برآورد پارامترهای مدل از روش کمترین توان‌های دوم به شیوه کپی و همکاران [۱۱] استفاده می‌کنیم. [۲] در این روش ملاک برآورد پارامترها، مینیمم سازی مجموع توان دوم فاصله بین توابع عضویت ترم‌های زبانی مشاهده شده و توابع عضویت ترم‌های زبانی پیش بینی شده توسط مدل است، یعنی

$$SSE = \sum_{i=۱}^{n} d^{۲}(\tilde{Y}_i, \tilde{y}_i)$$

که در آن $\tilde{y}_i$ مقدار تبدیل یافته (رجوع کنید به بخش ۱.۱.۲ ) مشاهده ترم‌های زبانی مربوط به هر ویژگی و $\tilde{Y}_i$ مقدار برآورد شده متناظر با $\tilde{y}_i$ است که در (۳.۴) تعریف شده است[۱]. همچنین $d^{۲}(\tilde{y}_i, \tilde{Y}_i)$ فاصله تعریف شده به شیوه کپی و دورسو [۱۱] است.

ایده اساسی در اینجا مدل‌سازی مراکز متغیر پاسخ فازی با استفاده از یک مدل رگرسیون خطی است. این ایده را می‌توان تحت شرایط زیر بیان کرد. ابتدا مدل مراکز مشاهده شده و مرزهای پایین و بالای متغیرهای پاسخ را با استفاده از مجموعه‌های مربوط به باقی‌مانده‌ها و مقادیر دیگر تعریف می‌کنیم[۱۱].

$$\mathbf{m} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

$$\mathbf{m} - \mathbf{l} = (\boldsymbol{\mu} - \delta_L) + \boldsymbol{\epsilon}_L$$

$$\mathbf{m} - \mathbf{u} = (\boldsymbol{\mu}\boldsymbol{\delta}_U) + \boldsymbol{\epsilon}_U.$$

در فرمول‌های بالا $\boldsymbol{\epsilon}$، $\boldsymbol{\epsilon}_L$ و $\boldsymbol{\epsilon}_U$ بردار خطاها و $\boldsymbol{\mu}$، $\boldsymbol{\delta}_L$ و $\boldsymbol{\delta}_U$ بردار مقادیر مراکز و پهناهای متغیر پاسخ است. بنابراین بردارهای معرفی شده براساس مدل رگرسیون به‌صورت زیر بررسی

می‌شوند.

$$\begin{aligned}\boldsymbol{\mu} &= X\boldsymbol{\beta}\\ {}_L &= \eta_L\boldsymbol{\mu} + \xi_L\mathbf{1}\\ \boldsymbol{\delta}_U &= \eta_U\boldsymbol{\mu} + \xi_U\mathbf{1}.\end{aligned} \tag{۱۲.۴}$$

ماتریس $\mathbf{X}$، ماتریس داده‌های مشاهده شده $x_{ij}$ها است، پارامترهای $\boldsymbol{\beta}$ و $(\eta_L, \eta_U, \xi_L, \xi_U)$ به‌ترتیب ضرایب خطی بین مراکز و توابع مربوط به $X_j$ و رابطه خطی بین پهنای مشاهده شده و مراکز برآورد شده هستند و 1 نشان دهنده برداری $(n \times ۱)$ با مقدار تکراری یک است.

روشی که برای برآورد پارمترهای فوق استفاده می‌شود مبتنی بر الگوریتم استاندارد است که شامل مینمم کردن فاصله بین مقادیر مشاهده شده متغیر $\tilde{y}_i$ و مقادیر برآورده شده $\tilde{Y}_i$ که به‌صورت $\tilde{Y}_i = (\boldsymbol{\mu}_i, \boldsymbol{\delta}_{L_i}, \boldsymbol{\delta}_{U_i}), i = ۱, \cdots, n$ تعریف می‌شود، مقادیر $\boldsymbol{\mu}_i$، $\delta_{L_i}$ و $\delta_{U_i}$ در مدل (۱۲.۴) تعریف شده است.

حال با توجه به معیار کمترین توان دوم، پارامترهای مربوط به (۱۲.۴) باید با به حداقل رساندن فاصله توان دوم بین مقادیر مشاهده شده متغیر پاسخ $\tilde{y}_i$ و مقادیر $\tilde{Y}_i$ برآورد شوند. برای این منظور از فاصله کپی و همکاران [۱۱] استفاده می‌کنیم.

$$\begin{aligned}d'^{۲}(\tilde{y}_i, \tilde{Y}_i) &= d'^{۲}\left((\mathbf{m}, \mathbf{l}, \mathbf{u})_{LR}, (\boldsymbol{\mu}, \boldsymbol{\delta}_L, \boldsymbol{\delta}_U)_{LR}\right)\\ &= \Delta^{۲}_{LR} = \|\mathbf{m} - \boldsymbol{\mu}\|^{۲} + \|(\mathbf{m} - \lambda\mathbf{l}) - (\boldsymbol{\mu} - \lambda\boldsymbol{\delta}_L)\|^{۲} + \|(\mathbf{m} + \rho\mathbf{u}) - (\boldsymbol{\mu} + \rho\boldsymbol{\delta}_U)\|^{۲}\\ &= ۳(\mathbf{m} - \boldsymbol{\mu})^\top(\mathbf{m} - \boldsymbol{\mu}) - ۲\lambda(m - \boldsymbol{\mu})^\top(\mathbf{l} - \boldsymbol{\delta}_L) + \lambda^{۲}(\mathbf{l} - \boldsymbol{\delta}_L)^\top(\mathbf{l} - \boldsymbol{\delta}_L)\\ &\quad + ۲\rho(\mathbf{m} - \boldsymbol{\mu})^\top(\mathbf{u} - \boldsymbol{\delta}_U) + \rho^{۲}(\mathbf{u} - \boldsymbol{\delta}_U)^\top(\mathbf{u} - \boldsymbol{\delta}_U)\end{aligned} \tag{۱۳.۴}$$

در این فاصله توابع وزن $\lambda$ و $\rho$ با توجه به تابع عضویت $LR$ به‌صورت $\lambda = \int_۰^۱ L^{-۱}(\omega)d\omega$ و $\rho = \int_۰^۱ U^{-۱}(\omega)d\omega$ تعریف می‌شوند، مقادیر مربوط به این توابع وزن در تنظیم مناسب پهنای چپ و راست در هنگام محاسبه مرزهای پایین و بالا اعداد فازی $\tilde{y}_i$ و $\tilde{Y}_i$ کمک می‌کند[۱۱].

در فاصله (۱۳.۴) می‌توانیم تابع کمترین توان دوم را براساس پارامترهای مدل (۱۲.۴)

$\boldsymbol{\beta}, \eta_L, \eta_U, \xi_L$ و $\xi_U$ جایگذاری کنیم :

$$
\begin{aligned}
&\min_{\boldsymbol{\beta},\eta_L,\eta_U,\xi_L,\xi_U} \Delta^{۲}_{LR}(\boldsymbol{\beta}, \eta_L, \eta_U, \xi_L, \xi_U) \\
&= ۳(\mathbf{m} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{m} - X\boldsymbol{\beta}) - ۲\lambda(\mathbf{m} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{l} - \mathbf{X}\boldsymbol{\beta}\eta_L - 1\xi_L) \\
&\quad + \lambda^{۲}(\mathbf{m} - \mathbf{X}\boldsymbol{\beta}\eta_L - 1\xi_L)^\top(\mathbf{l} - \mathbf{X}\boldsymbol{\beta}\eta_L - 1\xi_L) \\
&\quad + ۲\rho(\mathbf{m} - X\boldsymbol{\beta})^\top(\mathbf{u} - \mathbf{X}\boldsymbol{\beta}\eta_U - 1\xi_U) + \rho^{۲}(\mathbf{u} - \mathbf{X}\boldsymbol{\beta}\eta_U - 1\xi_U)^\top(\mathbf{u} - \mathbf{X}\boldsymbol{\beta}\eta_U - 1\xi_U) \\
&= ۳(\mathbf{m}^\top\mathbf{m} - ۲\mathbf{m}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top X\boldsymbol{\beta}) - ۲\lambda(\mathbf{m}^\top\mathbf{l} - \mathbf{m}^\top\mathbf{X}\boldsymbol{\beta}\eta_L - \mathbf{m}^\top 1\xi_L \\
&\quad + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{l} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\eta_L + \boldsymbol{\beta}^\top\mathbf{X}^\top 1\xi_L) \\
&\quad + \lambda^{۲}(\mathbf{l}^\top\mathbf{l} - ۲\mathbf{l}^\top X\boldsymbol{\beta}\eta_L - ۲\mathbf{l}^\top 1\xi_L + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\eta_L^{۲} + ۲\boldsymbol{\beta}^\top\mathbf{X}^\top 1\eta_L\xi_L + n\xi_L^{۲}) \\
&\quad + ۲\rho(\mathbf{m}^\top\mathbf{u} - \mathbf{m}^\top\mathbf{X}\boldsymbol{\beta}\eta_U - \mathbf{m}^\top 1\xi_U + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{u} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\eta_U + \boldsymbol{\beta}^\top\mathbf{X}^\top 1\xi_U) \\
&\quad + \rho^{۲}(\mathbf{u}^\top\mathbf{u} - ۲\mathbf{u}^\top\mathbf{X}\boldsymbol{\beta}\eta_U - ۲\mathbf{u}^\top 1\xi_U + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\eta_U^{۲} + ۲\boldsymbol{\beta}^\top\mathbf{X}^\top 1\eta_U\xi_U + n\xi_U^{۲}). \qquad (۱۴.۴)
\end{aligned}
$$

از تابع فوق نسبت به پارامترهای مجهول مشتق گرفته و مساوی صفر قرار می‌دهیم برآوردهای پارامترها به‌صورت زیر حاصل می‌شود.

$$
\begin{aligned}
\eta_L &= \lambda^{-1}(\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta})^{-1}\left[\lambda(\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{l} - \boldsymbol{\beta}^\top\mathbf{X}^\top 1\xi_L) - (\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{m} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta})\right], \\
\eta_U &= \rho^{-1}(\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta})^{-۱}\left[\rho(\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{u} - \boldsymbol{\beta}^\top\mathbf{X}^\top 1\xi_U) - (\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{m} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta})\right], \\
\xi_L &= (n\lambda)^{-۱}\left[\lambda 1^\top(\mathbf{l} - \mathbf{X}\boldsymbol{\beta}\eta_L) - 1^\top(\mathbf{m} - \mathbf{X}\boldsymbol{\beta})\right], \\
\xi_U &= (n\rho)^{-۱}\left[\rho 1^\top(\mathbf{u} - \mathbf{X}\boldsymbol{\beta}\eta_U) - 1^\top(\mathbf{m} - \mathbf{X}\boldsymbol{\beta})\right], \\
\boldsymbol{\beta} &= \left[۳ - \lambda\eta_L(۲ - \lambda\eta_L) + \rho\eta_U(۲ + \rho\eta_U)\right]^{-۱}(\mathbf{X}^\top\mathbf{X})^{-۱}\mathbf{X}^\top \\
&\quad \times \Big[۳\mathbf{m} - \lambda(\mathbf{m}\eta_L + \mathbf{l} + 1\xi_L) + \lambda^{۲}(\mathbf{l}\eta_L - 1\eta_L\xi_L) \\
&\qquad + \rho(\mathbf{m}\eta_U + \mathbf{u} + 1\xi_U) + \rho^{۲}(\mathbf{u}\eta_U - 1\eta_U\xi_U)\Big].
\end{aligned}
$$

باید به این نکته توجه داشت که برای به‌دست آوردن برآوردهای فوق، پیشنهاد می‌کنیم از یک الگوریتم تکراری با استفاده از چندین نقطه شروع مختلف برای ارزیابی پایداری جواب‌ها، استفاده کنید.

## ۱.۲.۴ مثال کاربردی

در داده‌های بیماری لوپوس (رجوع شود به بخش ۴.۲.۳ ) برای بدست آوردن ضرایب مدل با استفاده از روش پیشنهادی از نرم افزار Mathematica کمک گرفتیم. کدهای مربوط به آن را در

پیوست ب.۳ آورده‌ایم. مدل به دست آمده با استفاده از آنها به صورت زیر است.

$$\hat{\tilde{Y}}_i = (-۳/۹۹۶, ۰/۶۹۳, ۰/۷۳) + (-۰/۶۴۲, ۰, ۰/۰۲۸)ESRtest + ۰/۰۰۹AntiDNAtest$$
$$+ ۰/۰۴۴ANAtest + (۰/۱۲۴۰/۷۴۰, ۰)SunExpossure$$
$$+ (۰/۳۴۳, ۰/۲۵۸, ۰/۵۹۳)FamilyHistory$$

## ۲.۲.۴ معیار نیکویی برازش

در این بخش ضریب تعیین که یکی از معیارهای نیکویی برازش برای انتخاب متغیر است را تعریف می‌کنیم. برای این منظور ابتدا مفاهیم زیر را تعریف می‌کنیم.

**تعریف ۱.۲.۴.** برای متغیرهای خروجی LR می‌توان مفاهیم زیر را تعریف کرد:

- برای متغیر خروجی عدد فازی LR، مجموع مربعات کل از فاصله کپی و همکاران [۱۱] بین مقادیر مشاهده شده $\tilde{\mathbf{y}} = (\mathbf{m}, \mathbf{l}, \mathbf{u})_{LR}$ و میانگین مشاهدات $\bar{\tilde{y}} = (۱\bar{m}, ۱\bar{l}, ۱\bar{u})_{LR}$ به دست آید، به عبارت دیگر

$$d'^{۲}(\tilde{y}, \bar{\tilde{y}}) = SST = \|m - ۱\bar{m}\|^{۲} + \|(\mathbf{m} - \lambda\mathbf{l}) - (۱\bar{m} - \lambda۱\bar{l})\|^{۲} + \|(\mathbf{m} + \rho\mathbf{u}) - (۱\bar{m} + \rho۱\bar{u})\|^{۲}$$
$$= ۳(\mathbf{m} - ۱\bar{m}))^{\top}(\mathbf{m} - ۱\bar{m}) - ۲\lambda(\mathbf{m} - ۱\bar{m})^{\top}(\mathbf{l} - ۱\bar{l}) + \lambda^{۲}(\mathbf{l} - ۱\bar{l})^{\top}(\mathbf{l} - ۱\bar{l})$$
$$+ ۲\rho(\mathbf{m} - ۱\bar{m})^{\top}(\mathbf{u} - ۱\bar{u}) + \rho^{۲}(\mathbf{u} - ۱\bar{u})^{\top}(\mathbf{u} - ۱\bar{u})$$

که در آن $\bar{m}$، $\bar{l}$ و $\bar{u}$ میانگین مقادیر بردارهای $\mathbf{m}$، $\mathbf{l}$ و $\mathbf{u}$ هستند.

- مجموع مربعات رگرسیونی $(SSR)$، یعنی تغییراتی که توسط مدل مطرح شده اند، به صورت زیر تعریف می‌شوند.

$$d'^{۲}(\hat{\tilde{y}}, \bar{\tilde{y}}) = SSR = \|\hat{\mu} - ۱\bar{m}\|^{۲} + \|(\hat{\mu} - \lambda\hat{\delta}_L) - (۱\bar{m} - \lambda۱\bar{l})\|^{۲}$$
$$+ \|(\hat{\mu} + \rho\mathbf{u}) - (\bar{m} + \rho۱\hat{\delta}_U) - (۱\bar{m} + \rho۱\bar{u})\|^{۲}.$$

- مجموع مربعات خطا (باقیمانده) $(SSE)$، یعنی تغییراتی که توسط مدل مطرح نشده اند، به صورت زیر تعریف می‌شوند.

$$d'^{۲}(\tilde{y}, \hat{\tilde{y}}) = SSE = \|\mathbf{m} - \hat{\mu}\|^{۲} + \|(\mathbf{m} - \lambda\mathbf{l}) - (\hat{\mu} - \lambda\hat{\delta}_L)\|^{۲} + \|(\mathbf{m} + \rho\mathbf{u}) - (\hat{\mu} + \rho\hat{\delta}_R)\|^{۲}$$
$$= ۳(\mathbf{m} - \hat{\mu})^{\top}(\mathbf{m} - \hat{\mu}) - ۲\lambda(\mathbf{m} - \hat{\mu})^{\top}(\mathbf{l} - \hat{\delta}_L) + \lambda^{۲}(\mathbf{l} - \hat{\delta}_L)^{\top}(\mathbf{l} - \hat{\delta}_L)$$
$$+ ۲\rho(\mathbf{m} - \hat{\mu})^{\top}(\mathbf{u} - \hat{\delta}_U) + \rho^{۲}(\mathbf{u} - \hat{\delta}_U)^{\top}(\mathbf{u} - \hat{\delta}_U)$$

برای اندازه‌گیری نیکویی برازش مدل رگرسیون لجستیک با پاسخ فازی، از ضریب تعیین $R^{۲}$ و ضریب تعیین تعدیل یافته $R^{۲}_{adj}$ استفاده می‌کنیم. برای این منظور به تعریف آنها می‌پردازیم.

**تعریف ۲.۲.۴.** ضریب تعیین چندگانه [۱] به صورت زیر تعریف می‌شود.

$$R^۲ = ۱ - \frac{SSE}{SST} \qquad (۱۵.۴)$$

**تعریف ۳.۲.۴.** ضریب تعیین چندگانه تعمیم یافته [۲] $R^۲_{adj}$ به صورت زیر تعریف می‌شود.

$$R^۲_{adj} = ۱ - (۱ - R^۲)\frac{n-۱}{n-p-۴}. \qquad (۱۶.۴)$$

این شاخص شامل یک عامل تعدیل یافته بر مبنای تعداد ضرایب رگرسیون در مدل مرکزی، دو ضریب در مدل پهنای چپ و دو ضریب در مدل پهنای راست است.

## ۳.۲.۴ روش پیشرو برای انتخاب متغیر به شیوه کپی و همکاران

با توجه به تعریف‌هایی که قبل انجام دادیم، می‌توانیم تخمین یک مدل رگرسیونی مناسب را انجام دهیم. برای انتخاب مدل مناسب از الگوریتم زیر استفاده می‌کنیم: (شکل ۱.۴ )

۱. ابتدا مدل‌های تک متغیره می‌سازیم، از بین آنها مدلی که کمترین MPE را دارد انتخاب کرده و متغیر وارد شده در مدل را حفظ می‌کنیم.

۲. با حفظ متغیر در مدل، از ما بقی متغیرهای باقیمانده، متغیر جدیدی به مدل اضافه کرده و MPE هر یک از آنها را محاسبه می‌کنیم. مجدداً هر یک از مدل‌های که MPE کمتری داشت انتخاب می‌شود. به این ترتیب دو متغیر در مدل انتخاب می‌شود. این کار را تا جایی ادامه می‌دهیم که همه متغیرها وارد مدل شده باشند.

۳. در هر مرحله بعد از انتخاب مدل با کمترین MPE، مقدار $R^۲$ و $R^۲_{adj}$ را انتخاب می‌کنیم. زمانی که همه متغیرها وارد مدل شدند، از میان مدل‌های با کمترین MPE، مدلی را به عنوان مدل مناسب انتخاب می‌کنیم که بیشترین مقدار $R^۲_{adj}$ را داشته باشد.

در داده‌های مربوط بیماری لوپوس، ۵ متغیر تبیینی برای پیش‌بینی متغیر پاسخ دودویی داریم. مدل با همه متغیرها (۱۳.۳) است. در انجا می‌خواهیم با توجه به الگوریتمی که توضیح دادیم، انتخاب مناسب‌ترین مدل به کمک متر دورسو را انجام دهیم.

بیشترین مقدار $R^۲_{adj}$ در مرحله ۴ است، این مقدار بیان می‌کند که بهترین مدل،مدلی است که در آن متغیرهای $(ESRtest)x_۱$، $(AntiDNAtest)x_۲$، $(ANAtest)x_۳$ و $(FamilyHistory)x_۵$ حضور داشته باشند. مدل مناسب با استفاده از این روش به صورت زیر است.

$$\hat{\tilde{Y}} = \tilde{b}_۰ + \tilde{b}_۱ ESRtest + \tilde{b}_۲ AntiDNAtest + \tilde{b}_۳ ANAtest + \tilde{b}_۴ FamilyHistory \qquad (۱۷.۴)$$

[۱] Coefficient of multiple determination

[۲] Adjusted coefficient of multiple determination

جدول ۱.۴: مراحل انتخاب مدل

| شماره | MPE | $R^2$ | $R^2_{adj}$ | متغیرهای تبیینی |
|---|---|---|---|---|
| ۱ | ۱.۲۴۵ | ۰.۸۱۰ | ۰.۷۰۵ | $x_3$ |
| ۲ | ۰.۶۳۹ | ۰.۹۰۳ | ۰.۸۳۰ | $x_3, x_2$ |
| ۳ | ۰.۵۳۳ | ۰.۹۱۹ | ۰.۸۳۸ | $x_3, x_2, x_1$ |
| ۴ | ۰.۴۵۶ | ۰.۹۳۰ | ۰.۸۳۹ | $x_3, x_2, x_1, x_5$ |
| ۵ | ۰.۴۴۷ | ۰.۹۳۲ | ۰.۸۱۰ | $x_3, x_2, x_1, x_5, x_6$ |

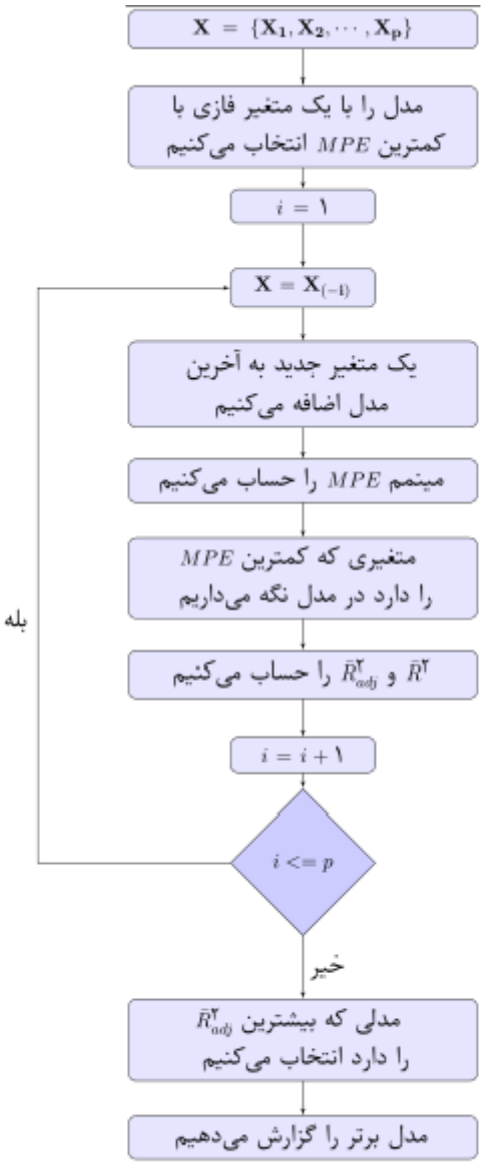
دلیل اینکه از ضریب تعیین تعدیل یافته برای انتخاب مدل مناسب انتخاب می‌کنیم این است که درصد نشان داده شده در ضریب تعیین تأثیر همه متغیرهای تبیینی بر متغیر وابسته است، به همین علت با اضافه کردن متغیر جدید به مدل مقدار ضریب تعیین افزایش می‌یابد. در حالی که درصد داده شده در ضریب تعیین تعدیل یافته فقط حاصل از تأثیر واقعی متغیرهای تبیینی مدل بر متغیر پاسخ است.

مدل (۱۷.۴) که با استفاده از روش انتخاب مدل متغیر کپی و همکاران [۱۱] انتخاب شده است، همانند مدل (۱۴.۳) که با شیوه سلمانی و همکاران [۲۷] با قرار دادن $LE = ۰/۷$، می‌باشد. به عبارت دیگر می‌توان گفت در امکان ابتلا به بیماری لوپوس عوامل $ANAtest$، $AntiDNAtest$ و $ESRtest$ و سابقه خانوادگی تاثیر گذار تر از سایر عوامل هستند.

مدل (۱۷.۴) را با استفاده از مدل پیشنهادی کپی و همکاران [۱۱] برازش می‌دهیم، در این صورت خواهیم داشت:

$$\begin{aligned}\hat{\hat{Y}} = {} & (-۴/۰۲۳, ۰/۷, ۰/۷۳۳)_T + (-۰/۶۱۴, ۰, ۰/۰۱۹, ۰/۰۱۹)_T ESRtest \\ & + ۰/۰۰۹ AntiDNAtest + (۰/۰۴۳۵, ۰/۰۰۰۹, ۰)_T ANAtest \\ & + (۰/۳۶۱, ۰/۲۳۷, ۰/۶۰۶)_T FamilyHistory \end{aligned} \quad (۱۸.۴)$$

مدل (۱۸.۴) نشان دهنده لگاریتم بخت امکانی فرد $i$ام به بیماری لوپوس است.




شکل ۱.۴: الگوریتم انتخاب متغیر کپی و همکاران در مدل رگرسیون لجستیک فازی

# بحث و نتیجه گیری و پیشنهادات برای آینده تحقیق

## بحث و نتیجه‌گیری

در این پایان‌نامه، مدلی که برای برآورد ضرایب با پاسخ دودویی ارائه کردیم، مدل رگرسیون لجستیک فازی است. رگرسیون لجستیک فازی همان رگرسیون لجستیک است که در فصل یک آن را معرفی کردیم. تفاوت آنها در این است که در رگرسیون لجستیک فازی متغیرهای تبیینی دقیق هستند اما متغیرهای پاسخ به جای اعداد صفر و یک به صورت مجموعه‌های فازی بین صفر و یک $[۰,۱]$ هستند. رگرسیون لجستیک فازی نیز با استفاده از تبدیل لجیت خطی می‌شود. در این مدل متغیر پاسخ و ضرایب فازی و متغیر تبیینی (پیشگو) دقیق هستند. دو روش استفاده شده در این پایان‌نامه برای برآورد ضرایب رگرسیون لجستیک فازی، روش امکانی و روش کمترین توان دوم خطا می‌باشد.

رگرسیون با رویکرد امکانی به دلیل اینکه توابع عضویت مجموعه‌های فازی اغلب به عنوان توزیع‌های امکانی توصیف می‌شوند، با این اسم نام‌گذاری شده‌اند. برای برآورد ضرایب فازی باید مدل فازی را به حداقل برسانیم، که یک مسئله برنامه‌ریزی خطی است و در آن تابع هدف مجموع پهناهای فازی است. برای این کار از روش‌های عددی استفاده می‌کنیم.

روش دیگر که برای برآورد ضرایب استفاده می‌شود، روش کمترین توان دوم خطا است. در این روش فاصله بین مقادیر مشاهده شده و مقادیر برآورد شده حداقل می‌شود. از دو فاصله متفاوت برای این منظور استفاده شد. فاصله که توسط ژو و همکاران [۳۵] معرفی شده است، یکی از این فواصل است. این فاصله در تحقیقات زیادی استفاده شده است. علاوه بر این چون این فاصله براساس $\alpha-$ برش‌ها به دست می‌آید حایز اهمیت است. همچنین تابع وزنی که در تعریف این فاصله وجود دارد، باعث انعطاف‌پذیری بیشتر این فاصله شده است.

از فاصله کپی و دورسو [۱۱] به عنوان یک روش پیشنهادی برای برآورد ضرایب مدل با رویکرد کمترین توان دوم استفاده شده است. برای حداقل کردن این فاصله، مراکز متغیر پاسخ فازی با استفاده از یک مدل رگرسیون خطی مدل‌سازی شدند. در این فاصله نیز مانند فاصله ژو [۳۵] ،از توابع وزن استفاده می‌شود که به تنظیم مناسب پهنای چپ و راست در

محاسبه فاصله بین $\tilde{y}$ و $\tilde{Y}$ کمک می‌کند.

در هر دو شیوه امکانی و کمترین توان دوم، مناسب‌ترین مدل را انتخاب کردیم. در این روش از میانگین توان دوم خطا و ضریب تعیین تعدیل یافته و معیار LE استفاده می‌شود. با انتخاب $LE = ۰/۷$ مدل انتخابی به روش امکانی و کمترین توان دوم از نظر انتخاب متغیر، یکسان شدند.

در نهایت باید گفت دلیل اصلی انجام این پایان‌نامه با موضوع مذکور، ابهام موجود در یافته‌های پزشکی است زیرا برای مدل‌سازی روابط بین داده‌های پزشکی، فرضیات آماری کافی نیستند. مجموعه‌های فازی می‌توانند جواب مفیدی برای ابهامات پزشکی راجع به امکان ابتلای یک فرد مشکوک به بیماری باشد.

# پیشنهادات

۱. بررسی رگرسیون لجستیک، با متغیرهای تبیینی فازی و متغیر پاسخ و ضرایب دقیق

۲. بررسی رگرسیون لجستیک، با متغیر تبیینی، متغیر پاسخ و ضرایب فازی

۳. تشخیص داده‌های پرت در رگرسیون لجستیک فازی

۴. انجام آزمون فرض‌های معنی داری پارامترها برای رگرسیون لجستیک فازی

# مراجع


[۱] پوراحمد، س. آیت‌اللهی ،س.م.ت. طاهری، س.م. حبیب آگهی، ز (۱۳۹۰). مدل‌سازی موقعیت‌های مبهم تشخیص در پزشکی با استفاده از رگرسیون لجستیک به روش کمترین مربعات فازی، **سری سیستم‌های فازس و محاسبات نرم**، جلد۲، ص ۱۹

[۲] خراباتی،میترا. ربیعی،محمدرضا،(۱۳۹۸). شیوه ای جدید در تحلیل رگرسیون لجستیک فازی، **نهمین سمینار آمار و احتمال فازی،بابلسر**

[۳] ربیعی،محمدرضا. خراباتی، میترا (۱۳۹۸). معرفی بسته fuzzyreg برای تحلیل برخی از روشهای رگرسیون فازی در R، **نهمین سمینار آمار و احتمال فازی،بابلسر**

[۴] سلمانی ف،(۱۳۹۵)، پایان‌نامه دکتری: مدل مارکف فازی وابسته به پیشگو و کاربرد آن در تحلیل پاسخ‌های مبهم طولی، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهید بهشتی

[۵] طاهری م، (۱۳۷۵)، **آشنایی با نظریه مجموعه‌های فازی** انتشارات جهاد دانشگاهی مشهد

[۶] طاهری م، ماشین چی م،(۱۳۹۲)، **مقدمه‌ای بر احتمال و آمار فازی** چاپ دوم، انتشارات دانشگاه شهید باهنر کرمان

[۷] مشکانی علی، (۱۳۸۳)، **تحلیل آماری داده‌های رسته‌ای** چاپ سوم، انتشارات دانشگاه فردوسی مشهد

[۸] میرزایی یگانه ش، ارقامی ن،(۱۳۸۶)، رگرسیون فازی: مروری بر چند رویکرد، **اندیشه آماری** ۴۷-۳۵:۲۳

[9] Akaike H.,(1974), A new look at the statistical model identification, **IEEE Trans Auto Control** 19, (6), pp 716

[10] Clemins A.,(1987), Least squares mthod fitting to fuzzy vectr data, **Fuzzy Sets and Systems** 22, pp 260-269

[11] Coppi A., Dourso P., Giordani P. and Santaro A.,(2006), Leasst squares estimatin of a linear regression model with LR fuzzy response, **Comp. Stat. Data. Anal** 51, pp 267-186

[12] Diamond P., (1987), Least squares method fitting of several fuzzy variable, Proc of second IFSA congres, Tokyo, pp 20

[13] Dubois D., Kerre E., Mesiar R., and Prade H., (2000), **Fuzzy interval analysis. In Fundamentals of fuzzy sets**, pp 483-581

[14] Gagolewski M., and Caha J., (2019), A Guide to the FuzzyNumbers Package for R.

[15] Gildeh B., Sand Gien D. E. N. I. S., (2002), A goodness of fit index to reliability analysis in fuzzy model, **3rd WSEAS international conference on fuzzy sets and fuzzy systems. Iterlaken, Switzerland**, pp 11-14

[16] Diamond P. and Korner R., (1997), Extaded fuzzy linear models and least squares estimated, **Comput. Math. Appl** 33, (9), pp 15-32

[17] Dubias D., (1980), **Fuzzy set system: Theory and Application** , Academic Press, New York

[18] Fauci A., Braunwald E., Kasper D., Hauser S., Longo D., Jameson J. and Loscalzo J., (2008), **Harrison's Principles of Internal Medicine** val:II, J.Weily, New York

[19] Hojati M., Bector C. R. and Smimou K., (2005), A simple method for computition of fuzzy linear regression, **Eyropean Journal of Operational Research** 166 , pp 172-182

[20] Kao C. and Chyu C. L., (2002), A fuzzy linear regression model with explantory power, **Fuzzy Sets and Systems** 126, pp 401-409

[21] Lee E. S. and Chang P. T., (1994), Fuzzy linear regression analysis with spread unconstraines in sign, **Comp. Math. Appl** 28, (4), pp 61-79

[22] Meye J. P., Stanley D. J., Herscovitch, L. and Topolnytsky, L., (2002), Affective, Continuance, and Normative Commitment to the Organization: A Metaanalysis of Antecedents, Correlates, and Consequences, **Journal of Vocational Behavior**, 61, pp 20–52

[23] McCullagh, P. and Nelder, J. A., (1989), **Generalized Linear Models**, 2nd ed, Chapman and Hall, London

[24] Mohammadi j. Taheri S. M., (2004), Pedomodels fitting with fuzzy least squares regression, **Iranian Journal of Fuzzy System**, (2), pp 45-61

[25] Nelder, J. A., Wedderburn, R. W. M., (1972), Generalized Linear Models, **Journal of the Royal Statistical Society, A** , 135 pp 370–384.

[26] Neter, J., Kutner M. H., Nachtsheim C. J. and Wasserman W., (1996), **Applied Linear Statistical Models**, Chicago: Irwin

[27] Salmani F., Taheri S. M. and Abad, A., (2019), A forward variable selection method for fuzzy logistic regression, **International Journal of Fuzzy Systems, 21(4)**, pp 1259-1269.

[28] Savic D.A. and Pedrycz W., (1991), Evaluation of fuzzy linear regression models, **Fuzzy Sets and Systems** 47, (2), pp 173-181

[29] Tabaei B. P. and Herman W. H., (2002), A multivariate logistic regression equation to screen for diabetes: development and validation, **Diabetes Care** 25(11), pp 1999-2003

[30] Tanaka H., Hayashi I. and Watada J., (1989), Possibilitic linear regression for fuzzy data, **Journal of Operational Research** 40, pp 389-396

[31] Tanaka H. and Ishibuchi H., (1991), Identifiction of possibilitic linear systems by quadratic membership function of fuzzy parameters, **Fuzzy Sets and Systems** 41, pp 145-160

[32] Pourahmad S., Ayatollahi S. M. T., Taheri S. M. and Agahi Z. H., (2011), Fuzzy logistic regression based on the least squares approach with application in clinical studies. **Computers and Mathematics with Applications**, 62(9), pp 3353-3365.

[33] Tanaka H., vejima D. and Asia K., (1982), Linear regresion analysis with fuzzy model, **IEEE Trans, system Man cyberent** 12, pp 903-907

[34] Van Broekhoven E. and De Baets B., (2006), Fast and accurate center of gravity defuzzification of fuzzy system outputs defined on trapezoidal fuzzy partitions. **Fuzzy sets and systems**, 157(7), pp 904-918

[35] Xu R. Li C. (2001), Multicimentional leas-squares fitting with a fuzzy model **Fuzzy Sets and Systems** 119, pp 215-223

[36] Zadeh L.A., (1975) , The consept of linguistic variable and its application to approximate reasonig I , II and III, **Information Sciences**, (8) pp 199-249

[37] Zimmermann H.J., (2001), **Fuzzy Set Theory and Its Appliction**, Kluwer, Boston 4rd ed

# پیوست آ

# معرفی بسته FuzzyNumbers

نام کامل این بسته Tools to Deal with Fuzzy Numbers است. [۳] این بسته از قابلیت معرفی اعداد فازی به روش‌های گوناگون برخوردار است. برای نصب نرم افزار از دستور

```
install.packages('FuzzyNumbers')
```

استفاده می‌کنیم. با دستور

```
library(FuzzyNumbers)
```

بسته را فراخوانی می‌کنیم.

## عدد فازی LR

عدد فازی A یک زیر مجموعه از $\mathbb{R}$ با تابع عضویت زیر است:

$$\mu_A(x) = \begin{cases} ۰ & x < a_۱ \\ left(\frac{x-a_۱}{a_۲-a_۱}) & a_۱ \leq x < a_۲ \\ ۱ & a_۲ \leq x \leq a_۳ \\ right(\frac{x_a-۳}{a_۴-a_۳}) & \\ ۰ & a_۴ < x \end{cases} \quad (۱.آ)$$

که $a_۱, a_۲, a_۳, a_۴ \in \mathbb{R}$ و $a_۱ \leq a_۲ \leq a_۳ \leq a_۴$ است. تابع $left : [۰, ۱] \longrightarrow [۰, ۱]$ و $left(۰) \geq ۰, left(۱) \leq ۱$ و تابعی غیر نزولی و $right : [۰, ۱] \longrightarrow [۰, ۱]$ و $right(۰) \geq ۱, right(۱) \leq ۰$ و تابعی غیر صعودی است.

**مثال** : $A_۱$ یک عدد فازی ذوزنقه‌ای است، در نرم افزار به‌صورت زیر تعریف می‌شود:

```
A1 <- FuzzyNumber(1, 2, 4, 7,
left=function(x) x,
right=function(x) 1-x
)
```

دستورات زیر جزییات مربوط به عدد فازی را نمایش می‌دهند.

```
 print(A1)
Fuzzy number with:
   support=[1,7],
      core=[2,4].

class(A1)
## [1] "FuzzyNumber"
## attr(,"package")
## [1] "FuzzyNumbers"

A1
## Fuzzy number with:
## support=[1,7],
## core=[2,4].

 summary(A1)
     Length       Class        Mode
          1 FuzzyNumber          S4
```

## عدد فازی مثلثی و ذوزنقه‌ای

اعداد فازی مثلثی با دستور

```
TrapezoidalFuzzyNumber(a_1, a_2, a_3, a_4)
```

و اعداد فازی مثلثی با دستور

```
TriangularFuzzyNumbe(a_1, a_2, a_3)
```

نمایش داده می‌شوند تحت شرط $a_۱ < a_۲ < a_۳ < a_۴$.
نمایش‌های دیگری برای اعداد فازی ذوزنقه‌ای و مثلثی وجود دارند که در زیر به آنها اشاره شده است.

```
T1 <- TrapezoidalFuzzyNumber(a_1, a_2, a_3, a_4)
T2 <- TrapezoidalFuzzyNumber(a_1,a_2,a_2,a_3)
 # or T2 <- TriangularFuzzyNumber(a_1, a_2, a_3)
TrapezoidalFuzzyNumber(a_1, a_1, a_2, a_2) # crisp interval
as.TrapezoidalFuzzyNumber(c(a_1, a_2)) # is the same
TrapezoidalFuzzyNumber(a_1, a_1, a_1, a_1) # crisp real
as.TrapezoidalFuzzyNumber(a_1) # is the same

T1 <- TrapezoidalFuzzyNumber(1, 1.5, 4, 7)
```

## عدد فازی قطعه‌ای خطی

عدد فازی قطعه‌ای خطی را نیز می‌توان در بسته FuzzyNumbers تولید کرد، برای این منظور از دستور زیر استفاده می‌کنیم.

```
PiecewiseLinearFuzzyNumber(a1, a2, a3, a4, knot.n = 0,
  knot.alpha = numeric(0), knot.left = numeric(0),
  knot.right = numeric(0))
```

knot.n تعدا برش (گره)های قطعه کننده عدد فازی را مشخص می‌کند. knot.left ابتدای برش‌ها، knot.right انتهای برش‌ها و knot.alpha ارتفاع برش‌ها را مشخص می‌کند.

```
P1 <- PiecewiseLinearFuzzyNumber(1, 2, 3, 4,
knot.n=1, knot.alpha=0.25, knot.left=1.5, knot.right=3.25)
 P1
Piecewise linear fuzzy number with 1 knot(s),
   support=[1,4],
      core=[2,3].
```

```
P2 <- PiecewiseLinearFuzzyNumber(1, 2, 3, 4,
knot.n=2, knot.alpha=c(0.25,0.6),
knot.left=c(1.5,1.8), knot.right=c(3.25, 3.5))
P2
Piecewise linear fuzzy number with 2 knot(s),
    support=[1,4],
     core=[2,3].
```

## اعمال تابع یکنوا بر عدد فازی

به منظور اعمال یک تابع یکنوا بر اعداد فازی از دستور fapply استفاده می‌شود. نکته قابل توجه این است که عدد فازی در اینجا باید از نوع قطعه‌ای باشد. در مثال زیر تابع $\sqrt{logA}$ اعمال شده است.

```
A <- as.PiecewiseLinearFuzzyNumber(
TrapezoidalFuzzyNumber(0,1,2,3), knot.n=100)
plot( fapply(A, function(x) sqrt(log(x+1))))
```

## محاسبه هسته، دامنه و برش‌های عدد فازی

هسته، دامنه و برش‌های یک عدد فازی به ترتیب با دستورهای supp ،core و alphacut قابل محاسبه‌اند.

```
A <- FuzzyNumber(-5, 3, 6, 20,
left=function(x) pbeta(x,0.4,3),
right=function(x) 1-x^(1/4),
lower=function(alpha) qbeta(alpha,0.4,3),
upper=function(alpha) (1-alpha)^4
)

supp(A)
[1] -5 20

core(A)
[1] 3 6
```

```
alphacut(A, 0)
L U
0 -5 20
```

## عملگرهای حسابی

عملگرهای دوبعدی [۱] ریاضی برای اعداد فازی متقارن برای هر $\alpha \in [۰, ۱]$ به‌صورت زیر تعریف می‌شوند:

$$(A \circledast B)_\alpha = A_\alpha \circledast B_\alpha$$

که $\circledast = +, -, \times, /$ و $A$ و $B$ اعداد فازی متقان هستند. به‌طور مثال جمع $A + B$ به‌صورت زیر تعریف می‌شود:

$$(A + B)_\alpha = A_\alpha + B_\alpha = [A_L(\alpha) + B_L(\alpha), A_U(\alpha) + B_U(\alpha)]$$

همچنین برای $\lambda \in \mathbb{R}$ ضرب اسکالر به‌صورت زیر تعریف می‌شود:

$$(\lambda.A)_\alpha = \alpha A_\alpha = \begin{cases} [\lambda A_L(\alpha), \lambda A_U(\alpha)], \lambda \geq ۰ \\ [\lambda A_U(\alpha), \lambda A_L(\alpha)], \lambda < ۰ \end{cases}$$

```
A <- TrapezoidalFuzzyNumber(0, 1, 1, 2)
B <- TrapezoidalFuzzyNumber(1, 2, 2, 3)
A+B
Trapezoidal fuzzy number with:
   support=[1,5],
      core=[3,3].
A-B
Trapezoidal fuzzy number with:
   support=[-3,1],
      core=[-1,-1].
3*A
Trapezoidal fuzzy number with:
   support=[0,6],
      core=[3,3].
```

[۱] binary

## تخمین قطعه‌ای یک عدد فازی

در بسته FuzzyNumbers، ۳ روش مختلف برای تخمین قطعه‌ای یک عدد فازی وجود دارد:

۱. کمترین فاصله اقلیدسی (NearestEuclidean)

۲. محافظ هسته و دامنه (SupportCorePreserving)

۳. ساده (Naive)

برای تخمین قطعه‌ای اعداد فازی از دستور زیر در نرم افزار استفاده می‌کنیم:

```
piecewiseLinearApproximation(object,
   method=c("NearestEuclidean", "SupportCorePreserving",
   "Naive"),
   knot.n=1, knot.alpha=seq(0, 1, length.out=knot.n+2)[-c(1,knot.n+2)],
   ..., verbose=FALSE)
```

حال به اختصار به تعریف روش‌ها می‌پردازیم.

۱. کمترین فاصله اقلیدسی در این روش، تخمین یک عدد فزی به وسیله یک عدد فازی قطعه‌ای به گونه‌ای انجام می‌گیرد که کمترین فاصله اقلیدسی را از عدد فازی اصلی داشته باشد.

۲. محافظ هسته و دامنه هسته و دامنه تخمین قطعه‌ای یک عدد فازی به این روش، معادل هسته و دامنه همان عدد فازی اصلی است.

۳. ساده در این روش، تخمین قطعه‌ای یک عدد فازی به گونه‌ای صورت می‌کیزد که:

(آ) هسته عدد فازی اصلی برار با هسته عدد فازی تخمینی

(ب) دامنه عدد فازی اصلی برابر با دامنه عدد فازی تخمینی

(ج) برش‌های معرفی شده در دستور برای عدد فازی اصلی، برابر با برش‌های متناظر عددهای فازی تخمینی

```
A <- FuzzyNumber(-5, 3, 6, 20,
left=function(x) pbeta(x,0.4,3),
right=function(x) 1-x^(1/4),
lower=function(alpha) qbeta(alpha,0.4,3),
upper=function(alpha) (1-alpha)^4
```

```
)

 (T1 <- trapezoidalApproximation(A, method='Naive'))
Trapezoidal fuzzy number with:
   support=[-5,20],
      core=[3,6].

(T2 <- trapezoidalApproximation(A, method='NearestEuclidean'))
Trapezoidal fuzzy number with:
   support=[-5.85235,14.4],
      core=[-2.26529,3.2].

 (T3 <- trapezoidalApproximation(A, method='ExpectedIntervalPreserving'))
Trapezoidal fuzzy number with:
   support=[-5.85235,14.4],
      core=[-2.26529,3.2].
```

## مینیمم و ماکزیمم عدد فازی

مینیمم عدد فازی برای برش‌ها به‌صورت

$$A_\alpha \wedge B_\alpha = [A_L(\alpha) \wedge B_L(\alpha), A_U(\alpha) \wedge B_U(\alpha)]$$

و ماکزیمم عدد فازی به‌صورت

$$A_\alpha \vee B_\alpha = [A_L(\alpha) \vee B_L(\alpha), A_U(\alpha) \vee B_U(\alpha)]$$

تعریف می‌شود.

مقدار مینیمم و ماکزیمم عدد فازی به‌صورت زیر قابل استفاده است.تعریف زیر فقط برای اعداد فازی خطی قطعه‌ای تعریف شده است.

```
x = as.PiecewiseLinearFuzzyNumber(TriangularFuzzyNumber
(-4.8, -3 , -1.5), knot.n = 9)
y = as.PiecewiseLinearFuzzyNumber(TriangularFuzzyNumber
(-5.5, -2.5, -1.1), knot.n = 9)
min = min(x@a1,y@a1)
max = max(x@a4,y@a4)
```

# پیوست ب

# کدهای استفاده شده در پایان‌نامه

## ب.۱ کدهای انتخاب مدل با معیارر LE

### کدهای R

```
library(FuzzyNumbers)
verylow<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(0.01,0.02,0.02,0.18), knot.n=100)
y1<-fapply(verylow, function(x) log(x/(1-x)))
T1 <- trapezoidalApproximation(y1, method='ExpectedIntervalPreserving')
Y1<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(supp(T1)[1],core(T1)[1],core(T1)[1],supp(T1)[2]), knot.n=100)
Y11=c(core(Y1)[1],core(Y1)[1]-supp(Y1)[1],supp(Y1)[2]-core(Y1)[2])
low<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(0.1, 0.25, 0.25, 0.4), knot.n=100)
y2<-fapply(low, function(x) log(x/(1-x)))
T2 <- trapezoidalApproximation(y2, method='ExpectedIntervalPreserving')
Y2<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
```

```
(supp(T2)[1],core(T2)[1],core(T2)[1],supp(T2)[2]), knot.n=100)
Y22=c(core(Y2)[1],core(Y2)[1]-supp(Y2)[1],supp(Y2)[2]-core(Y2)[2])
medium<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(0.35, 0.5, 0.5, 0.65), knot.n=100)
y3<-fapply(medium, function(x) log(x/(1-x)))
T3 <- trapezoidalApproximation(y3, method='ExpectedIntervalPreserving')
Y3<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(supp(T3)[1],core(T3)[1],core(T3)[1],supp(T3)[2]), knot.n=100)
Y33=c(core(Y3)[1],core(Y3)[1]-supp(Y3)[1],supp(Y3)[2]-core(Y3)[2])
high<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(0.6 , 0.75, 0.75, 0.9), knot.n=100)
y4<-fapply(high, function(x) log(x/(1-x)))
T4 <- trapezoidalApproximation(y4, method='ExpectedIntervalPreserving')
Y4<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(supp(T4)[1],core(T4)[1],core(T4)[1],supp(T4)[2]), knot.n=100)
Y44=c(core(Y4)[1],core(Y4)[1]-supp(Y4)[1],supp(Y4)[2]-core(Y4)[2])
veryhigh<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(0.8, 0.98, 0.98, 0.99), knot.n=100)
y5<-fapply(veryhigh, function(x) log(x/(1-x)))
T5 <- trapezoidalApproximation(y5, method='ExpectedIntervalPreserving')
Y5<- as.PiecewiseLinearFuzzyNumber(TrapezoidalFuzzyNumber
(supp(T5)[1],core(T5)[1],core(T5)[1],supp(T5)[2]), knot.n=100)
Y55=c(core(Y5)[1],core(Y5)[1]-supp(Y5)[1],supp(Y5)[2]-core(Y5)[2])
Y=rbind(Y44,Y33,Y44,Y44,Y33,Y55,Y33,Y44,Y22,Y11,Y22,Y33,Y22,Y22,Y33)
```

## کدهای Mathematica

```
Dade = Import["F:\\lopos.xlsx", {"Data", 1}];
n = Length[Transpose[Dade][[1]]];
p = Length[Dade[[1]]] - 3;
Y = Dade[[1 ;;, {p + 1, p + 2, p + 3}]];
Ybar = Mean[Y];
X = Dade[[1 ;;, Table[i, {i, p}]]];
t = Table[i, {i, 2, p}];
```

```
LE = 0.07;


MSE = Array[mse, p - 1];
For[i = 1, i <= p - 1, i++,
  X = Dade[[1 ;;, {1, i + 1}]];
  j = Length[Transpose[X]];
  A = Array[\[Beta], j, 0];
  Al = Array[\[Sigma]l, j, 0];
  Au = Array[\[Sigma]u, j, 0];
  Yhat = Array[z, {n, 3}];
     For[h = 1, h <= n, h++,
   Yhat[[h, 1]] = X[[h]].A;
   Yhat[[h, 2]] = X[[h]].Al;
   Yhat[[h, 3]] = X[[h]].Au;
   ];
  d = {};
  For[k = 1, k <= j, k++,
   d = Join[d, {Al[[k]] >= 0, Au[[k]] >= 0}]
   ];
  B = Join[A, Al, Au];
  \[Lambda] = 0.5;
  \[Rho] = 0.5;
  Q1 = NMinimize[{3*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 1]] -
          Yhat[[All, 1]]) -
      2 \[Lambda]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 2]] -
          Yhat[[All, 2]]) + \[Lambda]^2*(Y[[All, 2]] -
          Yhat[[All, 2]]).(Y[[All, 2]] - Yhat[[All, 2]]) +
      2 \[Rho]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 3]] -
          Yhat[[All, 3]]) + \[Lambda]^2*(Y[[All, 3]] -
          Yhat[[All, 3]]).(Y[[All, 3]] - Yhat[[All, 3]]), d} , B];
  MSE[[i]] = Q1[[1]]];
MSE
a1 = Min[MSE]
var1 = Which[a1 == MSE[[1]], t[[1]], a1 == MSE[[2]], t[[2]],
```

```
  a1 == MSE[[3]], t[[3]], a1 == MSE[[4]], t[[4]], a1 == MSE[[5]], t[[5]]]


 MSE = Array[mse, p - 2];
bb = Delete[t, var1 - 1]
For[i = 1, i <= p - 2, i++,
 X = Dade[[1 ;;, {1, var1, bb[[i]]}]];
 j = Length[Transpose[X]];
 A = Array[\[Beta], j, 0];
 Al = Array[\[Sigma]l, j, 0];
 Au = Array[\[Sigma]u, j, 0];
 Yhat = Array[z, {n, 3}];
    For[h = 1, h <= n, h++,
  Yhat[[h, 1]] = X[[h]].A;
  Yhat[[h, 2]] = X[[h]].Al;
  Yhat[[h, 3]] = X[[h]].Au;
  ];
 d = {};
 For[k = 1, k <= j, k++,
  d = Join[d, {Al[[k]] >= 0, Au[[k]] >= 0}]
  ];
 B = Join[A, Al, Au];
 \[Lambda] = 0.5;
 \[Rho] = 0.5;
 Q1 = NMinimize[{3*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 1]] -Yhat[[All, 1]]) -
     2 \[Lambda]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 2]] -
         Yhat[[All, 2]]) + \[Lambda]^2*(Y[[All, 2]] -
         Yhat[[All, 2]]).(Y[[All, 2]] - Yhat[[All, 2]]) +
     2 \[Rho]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 3]] -
         Yhat[[All, 3]]) + \[Lambda]^2*(Y[[All, 3]] -
         Yhat[[All, 3]]).(Y[[All, 3]] - Yhat[[All, 3]]), d} , B];
 MSE[[i]] = Q1[[1]];]
MSE
a2 = Min[MSE]
var2 = Which[a2 == MSE[[1]], bb[[1]], a2 == MSE[[2]], bb[[2]],a2 == MSE[[3]], bb[[3]],
```

```
 a2 == MSE[[4]],bb[[4]]]
 a2 - a1 > LE


 MSE = Array[mse, p - 3];
cc = Delete[t, {{var1 - 1}, {var2 - 1}}]
For[i = 1, i <= p - 3, i++,
 X = Dade[[1 ;;, {1, var1, var2, cc[[i]]}]];
 j = Length[Transpose[X]];
 A = Array[\[Beta], j, 0];
 Al = Array[\[Sigma]l, j, 0];
 Au = Array[\[Sigma]u, j, 0];
 Yhat = Array[z, {n, 3}];
    For[h = 1, h <= n, h++,
  Yhat[[h, 1]] = X[[h]].A;
  Yhat[[h, 2]] = X[[h]].Al;
  Yhat[[h, 3]] = X[[h]].Au;
  ];
 d = {};
 For[k = 1, k <= j, k++,
  d = Join[d, {Al[[k]] >= 0, Au[[k]] >= 0}]
  ];
 B = Join[A, Al, Au];
 \[Lambda] = 0.5;
 \[Rho] = 0.5;
 Q1 = NMinimize[{3*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 1]] -
         Yhat[[All, 1]]) -
     2 \[Lambda]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 2]] -
         Yhat[[All, 2]]) + \[Lambda]^2*(Y[[All, 2]] -
         Yhat[[All, 2]]).(Y[[All, 2]] - Yhat[[All, 2]]) +
     2 \[Rho]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 3]] -
         Yhat[[All, 3]]) + \[Lambda]^2*(Y[[All, 3]] -
         Yhat[[All, 3]]).(Y[[All, 3]] - Yhat[[All, 3]]), d} , B];
 MSE[[i]] = Q1[[1]];]
MSE
```

```
a3 = Min[MSE]
var3 = Which[a3 == MSE[[1]], cc[[1]], a3 == MSE[[2]], cc[[2]],
  a3 == MSE[[3]], cc[[3]]]
a3-a2>LE


MSE = Array[mse, p - 4];
dd = Delete[t, {{var1 - 1}, {var2 - 1}, {var3 - 1}}]
For[i = 1, i <= p - 4, i++,
 X = Dade[[1 ;;, {1, var1, var2, var3, dd[[i]]}]];
 j = Length[Transpose[X]];
 A = Array[\[Beta], j, 0];
 Al = Array[\[Sigma]l, j, 0];
 Au = Array[\[Sigma]u, j, 0];
 Yhat = Array[z, {n, 3}];
    For[h = 1, h <= n, h++,
  Yhat[[h, 1]] = X[[h]].A;
  Yhat[[h, 2]] = X[[h]].Al;
  Yhat[[h, 3]] = X[[h]].Au;
  ];
 d = {};
 For[k = 1, k <= j, k++,
  d = Join[d, {Al[[k]] >= 0, Au[[k]] >= 0}]
  ];
 B = Join[A, Al, Au];
 \[Lambda] = 0.5;
 \[Rho] = 0.5;
 Q1 = NMinimize[{3*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 1]] -
         Yhat[[All, 1]]) -
     2 \[Lambda]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 2]] -
         Yhat[[All, 2]]) + \[Lambda]^2*(Y[[All, 2]] -
         Yhat[[All, 2]]).(Y[[All, 2]] - Yhat[[All, 2]]) +
     2 \[Rho]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 3]] -
         Yhat[[All, 3]]) + \[Lambda]^2*(Y[[All, 3]] -
         Yhat[[All, 3]]).(Y[[All, 3]] - Yhat[[All, 3]]), d} , B];
```

```
 MSE[[i]] = Q1[[1]];]
MSE
a4 = Min[MSE]
var4 = Which[a4 == MSE[[1]], dd[[1]], a4 == MSE[[2]], dd[[2]]]
a4-a3>LE
```

# ب.۲ کدهای متر ژو

```
Dade = Import["F:\\lopos.xlsx", {"Data", 1}];
n = Length[Transpose[Dade][[1]]];
p = Length[Dade[[1]]] - 3;
X = Dade[[1 ;;, Table[i, {i, p}]]];
AA = Dade[[1 ;;, {p + 1, p + 2, p + 3}]];
A = Array[\[Beta], p, 0];
Sl = Array[\[Sigma]l, p, 0];
Sr = Array[\[Sigma]r, p, 0];
What = Array[z, {n, 3}];
 For[h = 1, h <= n, h++,
  What[[h, 1]] = X[[h]].A;
  What[[h, 2]] = X[[h]].Sl;
  What[[h, 3]] = X[[h]].Sr;
  ];
  dis = Array[q, {n, 2}];

For[i = 1, i <= n, i++,
  dis[[i, 1]] =
   Log[(AA[[i, 1]] - (AA[[i, 2]]*(1 - \[Alpha])))/(
    1 - AA[[i, 1]] + (AA[[i, 2]]*(1 - \[Alpha])))];
  dis[[i, 2]] =
   Log[(AA[[i, 1]] + (AA[[i, 3]]*(1 - \[Alpha])))/(
    1 - AA[[i, 1]] - (AA[[i, 3]]*(1 - \[Alpha])))]
  ];
  d = {};
For[k = 1, k <= p, k++,
```

```
  d = Join[d, {Sl[[k]] >= 0, Sr[[k]] >= 0}]
  ];
B = Join[A, Sl, Sr];
Z = \!\(
\*UnderoverscriptBox[\(\[Sum]\), \(t =
    1\), \(15\)]\((\[Alpha] . \((\((dis[\([t,
          1]\)] - \((\((\[Alpha] - 1)\)*What[\([t, 2]\)])\) -
        What[\([t, 1]\)])\)^2 + \((dis[\([t,
          2]\)] - \((\((1 - \[Alpha])\)*What[\([t, 3]\)])\) -
        What[\([t, 1]\)])\)^2)\))\)\)
        intt = \!\(
\*SubsuperscriptBox[\(\[Integral]\), \(0\), \(1\)]\(Z \
\[DifferentialD]\[Alpha]\)\)
NMinimize[{intt, d}, B, Method -> "DifferentialEvolution"];
```

# ب.۳ کدهای متر کپی و همکاران

## برآورد ضرایب

```
Dade = Import["F:\\lopos.xlsx", {"Data", 1}];
n = Length[Transpose[Dade][[1]]];
p = Length[Dade[[1]]] - 3;
Y1 = {{"MPE", "Beta"}};
Y = Dade[[1 ;;, {p + 1, p + 2, p + 3}]];

A = Array[\[Beta], p, 0];
Al = Array[\[Sigma]l, p, 0];
Au = Array[\[Sigma]u, p, 0];
Yhat = Array[z, {n, 3}];
X = Dade[[1 ;;, Table[i, {i, p}]]];
   For[i = 1, i <= n, i++,
  Yhat[[i, 1]] = X[[i]].A;
```

```
  Yhat[[i, 2]] = X[[i]].Al;
  Yhat[[i, 3]] = X[[i]].Au;
  ];
d = {};
For[k = 1, k <= p, k++,
  d = Join[d, {Al[[k]] >= 0, Au[[k]] >= 0}]
  ];
B = Join[A, Al, Au];
\[Lambda] = 0.5;
\[Rho] = 0.5;
Q1 = NMinimize[{3*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 1]] -
        Yhat[[All, 1]]) -
    2 \[Lambda]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 2]] -
        Yhat[[All, 2]]) + \[Lambda]^2*(Y[[All, 2]] -
        Yhat[[All, 2]]).(Y[[All, 2]] - Yhat[[All, 2]]) +
    2 \[Rho]*(Y[[All, 1]] - Yhat[[All, 1]]).(Y[[All, 3]] -
        Yhat[[All, 3]]) + \[Lambda]^2*(Y[[All, 3]] -
        Yhat[[All, 3]]).(Y[[All, 3]] - Yhat[[All, 3]]), d} , B]
```

**Aabstract**

Logistic regression helps in the modeling of discrete statistical data. These models are very useful in medical sciences (such as death / life, the presence of disease / non-disease, etc.). The in diagnosis, doctors use various information sources, such as clinical history, medical examinations, laboratory tests, and so on. But placing people in both the sick and healthy groups is basically ambiguous, and obscure observations in clinical diagnosis are abundant. In such cases, logistic regression is not appropriate due to the lack of accuracy of the response variable. For fuzzy binary fuzzy response modeling, the fuzzy logistic regression model is presented in this thesis. To calculate the success of a fuzzy logistic model, the term linguistic terminology such as ..., low, medium, high, ... is defined. Then, using the principle of expansion, the transformation of the logic "odds ratio" is modeled on a set of observations of precise explanatory variables. also to estimate the coefficients of the proposed model, we use the probabilistic method and the least squares fuzzy method. We also offer two model selection methods to evaluate the model.

Keywords : Fuzzy Regression, Logistic Regression, Fuzzy Logistic Regression, Least Squares Error, Model Selection


**Shahrood University of Technology**

**Faculty Of Mathematical Sciences**

**MSc Thesis in: Mathematical Statistics**

# Logistic Regression in Fuzzy Environment and its application in medical sciences

**By: Mitra Kharabati**

**Supervisor**

**Mohammad Reza Rabiee**

**Advisor**

**Fatemeh Salmani**

**September 2019**