Journal of Statistical Sciences, Autumn and Winter, 2024
Vol. 18, No. 2, pp 395-414
DOI: 10.61186/jss.18.2.6

Fuzzy Multiple Logistic Regression Under Fuzzy Random Errors

Maleki, M., Nilisani, H. R. Akbari, M. G.
Department of Statistics, University of Birjand, Birjand, Iran.

Corresponding author: H. R. Nili-Sani, hnilisani@birjand.ac.ir

Received: 27/1/2024 **Revised:** 26/9/2024 **Accepted and Published Online:** 27/9/2024.

Introduction

Data classification is one of the most critical subjects in machine learning and statistics. Several methods have been suggested for classification. One of the appropriate and efficient methods is logistic regression for crisp variables. But a big part of data in psychology, Sociology, clinical research, etc., is vague and random. In these cases, the classification is also vague, and the essential assumptions of logistic regression are invalid. In this paper, we generalized ordinary logistic regression to fuzzy random variables. We may consider response, error, or independent random variables or parameters non-precise. We tried to fit an appropriate logistic regression model for a case of fuzzy random variables and to determine the recurrent relations such that the possibility of updating the model was provided.

Material and Methods

It is a well-known fact that $\tilde{u}$ is a fuzzy number if and only if the 1-cut set isn't empty and for all $(0 \leq \alpha \leq 1)$, α-cut set is a closed and bounded interval (Goetschel and Voxman, 1986). An especial case of fuzzy numbers is L-R fuzzy numbers where L and R are non- increasing nonnegative real –valued function on $[0, \infty)$ (Taheri and Mashinchi, 2007). The set of fuzzy number is denoted by $F(\mathbb{R})$. The function $\tilde{X}^* : \Omega \to F(\mathbb{R})$ is called a fuzzy random variable, if is measurable (Joo and Kim, 2001). A triangular fuzzy random variable is denoted by $\tilde{X}^* = (X^L, X, X^U)$ if X^L, X and $X^U$are three random variables (Hesamian and Akbari, 2022).

Results and Discussion

In the remainder, Y is response random variable, $\mathbf{X}$ is a vector of independent

random variables, $\tilde{\epsilon}^*$ is fuzzy random variable and $\tilde{\mu} = E(Y|\mathbf{X}) = \boldsymbol{\beta}'\mathbf{X} \oplus \tilde{\epsilon}$.

Lemma 1: Under the conditions of the above model if $\tilde{\epsilon}$ is a fuzzy number, $\tilde{\epsilon} = (m^*; a, b)_{LR}$, with the support of $[0, 1]$, then $\tilde{w} = \ln(\frac{\tilde{\mu}}{1-\tilde{\mu}}) = \boldsymbol{\beta}'\mathbf{x} \oplus \tilde{\epsilon}$, is a fuzzy number with the following membership function

$$\tilde{w}(y) = \begin{cases} L(\frac{m - \frac{1}{1+e^{-y}}}{a}), & \frac{1}{1+e^{-y}} \leq m, \\ R(\frac{\frac{1}{1+e^{-y}} - m}{b}), & \frac{1}{1+e^{-y}} \geq m, \end{cases}$$

Theorem 1: Suppose $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are n independent observations of model $Y = \boldsymbol{\beta}'\mathbf{X} \oplus \tilde{\epsilon}^*$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$, and $\tilde{\epsilon}_i^*$ are triangular fuzzy random variables, $\tilde{\epsilon}_i^* = (e_i; X_i^L, X_i^U)_T$, where e_i is the center of error(=0) and X_i^L and X_i^U are uniform random variables on $(0, 1)$. The least-square estimate of $\boldsymbol{\beta}$ of $\tilde{w} = \ln(\frac{\tilde{\mu}}{1-\tilde{\mu}}) = \boldsymbol{\beta}'\mathbf{x} \oplus \tilde{\epsilon}$, and by Xu and Li's meter with $f(\alpha) = \alpha$ is given by $\boldsymbol{\beta} = 2(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i')^{-1} \sum_{i=1}^n z_i\mathbf{x}_i$, where

$$z_i = \int_0^1 \alpha(\ln \frac{h_{il}(\alpha - 1) + m_i}{1 - h_{il}(\alpha - 1) - m_i})d\alpha + \int_0^1 \alpha(\ln \frac{h_{ir}(1 - \alpha) + m_i}{1 - h_{ir}(1 - \alpha) - m_i})d\alpha.$$

Conclusion

In this study, we discussed logistic regression and assumed that the error and response are triangular fuzzy random variables, and the explanatory variables are crisp. Model parameters were estimated using the least squares approach, and a recurrence method for computing parameters was obtained. Two examples have been used to detect the applied aspect of our model.

Keywords: Multiple Logistic Regression, Fuzzy Random Variable.

Mathematics Subject Classification (2010): 62J12, 62H30.

مجله علوم آماری، پاییز و زمستان ۱۴۰۳
جلد ۱۸، شماره ۲، ص ۳۹۵ -- ۴۱۴
DOI: 10.61186/jss.18.2.6
مقاله پژوهشی

رگرسیون لوژستیک چندگانه فازی تحت خطاهای تصادفی فازی

مریم مالکی، حمیدرضا نیلی‌ثانی، محمد قاسم اکبری
گروه آمار، دانشگاه بیرجند

نویسنده مسئول : حمیدرضا نیلی‌ثانی، hnilisani@birjand.ac.ir

تاریخ دریافت: ۱۴۰۲/۱۱/۷ تاریخ بازنگری: ۱۴۰۳/۷/۵ تاریخ پذیرش و انتشار: ۱۴۰۳/۷/۶

چکیده: در این مقاله، موضوع طبقه‌بندی داده‌ها مدنظر قرار داده می‌شود که در آن متغیر پاسخ به‌صورت دو یا چند ارزشی و متغیرهای پیشگو متغیرهای معمولی هستند اما، خطاها علاوه بر ماهیتی تصادفی، ماهیتی ابهامی نیز دارند. در این صورت متغیر پاسخ نیز متغیر تصادفی فازی است. بر این اساس مدلی بر پایه رگرسیون لوژستیک صورت‌بندی کرده و برآورد ضرایب با استفاده از روش کمترین توان‌های دوم بدست آورده می‌شود. با یک مثال نتایج حاصله برای حالت یک متغیر مستقل تشریح می‌گردند. در پایان روابط بازگشتی برای محاسبه برآورد پارامترها ارائه می‌شوند. این روابط بازگشتی می‌توانند در یادگیری ماشین و برای طبقه‌بندی داده‌های بزرگ مورد استفاده قرار گیرند.

واژه‌های کلیدی: رگرسیون لوژستیک چندگانه، متغیر تصادفی فازی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J12 ، 62H30.

۱ مقدمه

طبقه‌بندی داده‌ها در آمار و یادگیری ماشین از اهمیت زیادی برخوردار است و روش‌های مختلفی برای آن پیشنهاد شده است. یکی از روش‌های مناسب و کارآمد طبقه‌بندی داده‌ها، رگرسیون لوژستیک است آگریستی (۲۰۱۵). از این روش برای رده‌بندی انواع متنوعی از داده‌ها استفاده می‌شود مقیم‌بیگی (۱۴۰۱). با گسترش فناوری اطلاعات امکان ذخیره‌سازی و پردازش مه داده‌ها، یعنی مجموعه داده‌هایی که اندازه آن فراتر از توانایی ابزارهای نرم افزارهای رایج برای پردازش و تحلیل داده است، از جمله داده‌های موجود در شبکه‌های اجتماعی، نجوم، ژنتیک و غیره، بیش

از گذشته فراهم‌شده است. بخش قابل‌توجهی از این داده‌ها علاوه بر ماهیت تصادفی، ماهیت ابهامی یا نادقیق نیز دارند. در بسیاری از تحقیقات مرسوم و معمول، از جمله در تحقیقات بالینی، علوم اجتماعی و روانشناسی نیز چنین داده‌هایی مشاهده می‌شوند. اما، چنانچه داده‌ها نادقیق یا فازی باشند، مرزهای طبقه‌بندی نیز مبهم و نادقیق خواهند بود. به عنوان نمونه چنانچه هدف طبقه‌بندی داده‌ها به دو رده متمایز باشند، به دلیل نادقیق بودن مشاهدات، این طبقه‌بندی نیز نادقیق است و در چنین مواردی فرض‌های مورد نیاز برای رگرسیون برقرار نخواهد بود. مثلاً، در رگرسیون لوژستیک معمولی متغیر پاسخ دو حالتی دارای توزیع برنولی است، اما اگر متغیر پاسخ نادقیق باشد، چنین فرضی در مورد متغیر پاسخ برقرار نیست. یک راهکار این است که مشاهدات نادقیق حذف شوند. این رویکرد غیر منطقی است، زیرا ممکن است همه مشاهدات نادقیق باشند. ناکارامدی رگرسیون لوژستیک معمولی برای مدل‌بندی مشاهدات رسته‌ای فازی و فراوانی اینگونه مشاهدات و ضرورت بررسی، مدل‌بندی و به خصوص طبقه‌بندی داده‌های نادقیق، از جنبه‌های مختلف و با روش‌های مختلف مورد توجه محققین قرار گرفته است.

تحلیل مدل‌های رگرسیونی لوژستیک، وقتی متغیرهای پیشگو و وابسته به‌صورت فازی گزارش شوند، توسط تاکمورا (۲۰۰۴) مورد مطالعه قرارگرفته است. تسینگ و لین (۲۰۰۵) برای پیش‌بینی ورشکستگی در مؤسسات مالی انگلستان از رگرسیون لوژستیک دو ارزشی بازه‌ای درجه دوم با ضرایب فازی استفاده کردند. آنها مسئله رگرسیونی را به عنوان یک مساله برنامه‌ریزی درجه دوم ریاضی مدل‌بندی کردند. در روش آنها، ابهام کلی بر مبنای پهنای باندها و مجموع توان دوم فاصله بین مراکز خروجی برآورد شده و مشاهدات ورودی نسبت به قیودی بر حسب الفا برش‌ها مینیمم می‌شوند. ناگار و سریواستاوا (۲۰۰۸) مدلی برای رگرسیون لوژستیک فازی معرفی و مدل را برای پیش‌بینی یک نوع سرطان موردبررسی قرار داده و نتایج را با شبکه‌های عصبی فازی مقایسه نمودند. دام (۲۰۰۹) به بررسی عوامل خطر مرتبط با سرطان دهان از طریق رگرسیون فازی پرداخت. بررسی رگرسیون لوژستیک دو ارزشی بازه‌ای و رگرسیون لوژستیک دو ارزشی امکانی توسط طاهری و میرزایی (۲۰۰۹) و میرزایی و طاهري (۲۰۱۰) انجام‌شده است. پوراحمد و همکاران (a،b۲۰۱۱) رگرسیون لوژستیک دو ارزشی را برای حالتی که ورودی‌ها اعدادی دقیق و خروجی‌ها و ضرایب اعداد فازی باشند، ارائه داده‌اند و به روش کمترین توان دوم فازی پارامترها را برآورد کردند. نامداری و همکاران (۲۰۱۳) مدل‌های رگرسیونی لوژستیک فازی را وقتی متغیرهای پیشگو به‌صورت معمولی و متغیرهای پاسخ به‌صورت متغیرهای زبانی بیان شوند، موردبررسی و مطالعه قرار داده‌اند. یو و تسینگ (۲۰۰۵) مدل‌های رگرسیونی لوژستیک فازی را برای پردازش نور در یک کارخانه تلویزیون به کار گرفتند. نامداری و همکاران (۲۰۱۵) مدل‌های رگرسیونی لوژستیک فازی را بر اساس روش کمترین قدر مطلق انحرافات بررسی کرده‌اند. حسامیان و اکبری (۲۰۱۷) به بررسی مدل‌های رگرسیون لوژستیک نیمه پارامتری با متغیر وابسته فازی شهودی و متغیرهای پیشگو دقیق پرداخته‌اند. سلمانی و همکاران (۲۰۱۷) مدل‌های رگرسیون لوژستیک، وقتی متغیرهای مستقل و وابسته به‌صورت فازی گزارش شوند، موردمطالعه قراردادند.

رده بندی یک روش یادگیری با ناظر است. روش‌های رده‌بندی را می‌توان به دو دسته کلی تقسیم‌بندی کرد، روش‌های مبتنی بر یادگیری آماری، مانند روش‌های بیزی و روش‌های رگرسیون لوژستیک و روش‌های مبتنی بر یادگیری ماشین، مانند روش‌های مبتنی بر درخت و روش‌های ماشین‌بردار پشتیبان، برخی از روش‌ها نیز از تلفیق

روش‌های یادگیری آماری و یادگیری ماشین ایجاد می‌شوند. در روش‌های مبتنی بر یادگیری ماشین، که عموماً برای داده‌های بزرگ استفاده می‌شوند، هیچ فرض آماری پذیرفته نمی‌شود. در مقابل روش یادگیری آماری بر فرضیات آماری و ساختار احتمالی نمونه‌های تصادفی بنا شده و از این فرضیات و ساختارها در رده‌بندی، برآورد و آزمون فرض‌ها استفاده می‌شود. در تحقیقاتی که در زمینه رگرسیون لوژستیک فازی تاکنون انجام‌شده است (تا آنجایی که مؤلفین اطلاع دارند) هیچ ساختار احتمالی برای داده‌ها یا خطاهای تصادفی که ناشی از انتخاب تصادفی نمونه‌ها یا اندازه‌گیری‌ها است، در نظر گرفته نشده‌اند یا خطاهای موجود در مدل با فازی بودن ضرایب یا فازی بودن متغیرها تلفیق شده‌اند. درصورتی‌که این خطاها در بسیاری از موارد مشهود و با ماهیت فازی متفاوت است. به‌عنوان مثال هنگامی‌که وجود یا عدم وجود درد در بیماران خاص، مثلاً تحت یک عمل جراحی، مورد بررسی قرار می‌گیرد، میزان دردی که توسط فرد بیمار گزارش می‌شود از اهمیت خاصی برخوردار است. میزان درد هرچند که ماهیتی ابهامی دارد، آستانه و میزان درد از فردی به فرد دیگر می‌تواند متفاوت باشد، بنابراین ماهیتی تصادفی نیز دارد. در صورت نادیده گرفتن ماهیت تصادفی داده‌ها، همواره این امکان وجود دارد که مدل خوبی به داده‌ها برازش شود. پوراحمد و همکاران (a۲۰۱۱) و نیکویی برازش با معیارهای مختلفی مانند قابلیت یا تعمیم مفاهیم مانند دقت اندازه‌گیری شود. اما، با توجه به اینکه در اکثر موارد اندازه نمونه کوچک است، در چنین شرایطی تضمینی وجود ندارد که مدل منتخب به‌خوبی به مجموعه دیگری از داده‌ها برازش گردد، چون مشخص نیست که ساختار احتمالی دو مجموعه داده‌ها یکسان باشند. از سوی دیگر، در صورت مواجهه با یک مجموعه از داده‌ها و مدل‌بندی آنها ضرورت دارد تا با تعیین روابط بازگشتی این امکان فراهم آید که با افزودن یک یا چند داده جدید، مدل به سهولت اصلاح گردد. در این مقاله تلاش می‌شود با ورود جمله خطا به مدل که ماهیتی تصادفی و فازی دارد، مدل مناسب به داده‌ها برازش و با تعیین روابط بازگشتی امکان به‌روزرسانی سریع مدل فراهم آید. در بخش ۲، مفاهیم مقدماتی که در بخش‌های بعدی مورد استفاده و استناد قرار خواهند گرفت، ارائه می‌شوند. بخش ۳به ارائه نتایج اصلی اختصاص داده شده است. در بخش ۴ و به منظور تشریح نتایج اصلی یک مثال ارائه خواهد شد. جمع‌بندی و نتیجه گیری در بخش ۵ آورده شده است.

۲ مفاهیم مقدماتی

تعریف ۱. تابع $\tilde{u} : \mathbb{R}^n \to [0,1]$ یک مجموعه فازی از مجموعه مرجع $\mathbb{R}^n$ نامیده می‌شود و α–برش مجموعه فازی $\tilde{u}$ به صورت

$$\tilde{u}_\alpha = \begin{cases} \{x \in \mathbb{R}^n : \tilde{u}(x) \geq \alpha\}, & 0 < \alpha \leq 1, \\ \text{supp } \tilde{u}, & \alpha = 0, \end{cases}$$

تعریف می‌شود، که در آن $\text{supp } \tilde{u} = cl\{x \in \mathbb{R}^n : \tilde{u}(x) > 0\}$. مجموعه فازی $\tilde{u}$ را نرمال نامند، اگر $\tilde{u}_1 \neq \varnothing$ و محدب نامند، اگر برای هر $0 \leq \alpha \leq 1$ ، $\tilde{u}_\alpha$ مجموعه محدب باشد. $\tilde{u}$ را تابع عضویت نیز می‌نامند.

تعریف ۲. مجموعه فازی $\tilde{u}$ را بر $\mathbb{R}$ یک عدد فازی گویند، اگر $\tilde{u}$ مجموعه فازی محدب، نرمال، نیم‌پیوسته بالایی و $\text{supp}\ \tilde{u}$ فشرده باشد. خانواده همه اعداد فازی بر $\mathbb{R}$ با $F(\mathbb{R})$ نشان داده می‌شود.

بنابراین در حالتی که $\tilde{u}$ بر $\mathbb{R}$ تعریف شده باشد، $\tilde{u}$ یک عدد فازی است اگر و تنها اگر $\tilde{u}_1 \neq \varnothing$ و $\tilde{u}_\alpha$ برای هر $0 \leq \alpha \leq 1$، یک فاصله بسته و کراندار، $\tilde{u}_\alpha = [\tilde{u}_{\ell\alpha}, \tilde{u}_{r\alpha}]$ باشد. خواص بیشتری از اعداد فازی $\tilde{u} \in F(\mathbb{R})$ را می‌توانید در گوتشل و واکسمن (۱۹۸۶) بیابید. در ادامه فرض می‌شود $u_\ell(\alpha) = u_{\ell\alpha}$ و $u_r(\alpha) = u_{r\alpha}$ ، $0 \leqslant \alpha \leqslant 1$.

فرض کنید L و R توابع غیر صعودی از $\mathbb{R}^+$ به $[0,1]$ باشند و $L(0) = R(0) = 1$. اگر ساختار تابع عضویت عدد فازی $\tilde{u}$ به‌صورت

$$\tilde{u}(x) = \begin{cases} L(\frac{m-x}{a}), & x \leq m, \\ R(\frac{x-m}{b}), & x \geq m, \end{cases}$$

باشد، $\tilde{u}$ را یک عدد فازی LR نامیده و با نماد $(m; a, b)_{LR}$ نشان داده می‌شود. عدد حقیقی m را مرکز و اعداد مثبت a و b به ترتیب پهنای چپ و راست می‌نامند. اگر $L = R$ و $a = b$ آنگاه عدد فازی را متقارن و اگر $L = R = \max\{0, 1 - |x|\}$ آن را عدد فازی مثلثی نامند و با نماد $(m; a, b)_T$ نمایش می‌دهند.

تعریف ۳. اصل گسترش، (طاهری و ماشین چی ، ۱۳۸۷) فرض کنید $X_1, \ldots, X_n$، n مجموعه مرجع، $X = X_1 \times \ldots \times X_n$ حاصلضرب دکارتی آنها، $A_1, \ldots, A_n$، n مجموعه فازی به ترتیب از $X_1, \ldots, X_n$ و $y = f(x_1, \ldots, x_n)$ یک نگاشت از X به Y باشند. حاصل عمل f بر n مجموعه فازی A_n ، ...، A_1 به صورت مجموعه فازی B از Y با تابع عضویت

$$B(y) = f(A_1, \ldots, A_n)(y) = \begin{cases} \sup\limits_{X_1,\ldots,X_n, y=f(X_1,\ldots,X_n)} \min\{A_1(x_1), \ldots, A_n(x_n)\}, & f^{-1}(y) \neq \varnothing, \\ 0, & f^{-1}(y) \neq \varnothing, \end{cases}$$

تعریف می‌شود، که در آن $f^{-1}(y)$ نگاشت معکوس y تحت f است.

فرض کنید $\tilde{\mu} = (m_1; a_1, b_1)_T$ و $\tilde{\nu} = (m_2; a_2, b_2)_T$ دو عدد فازی مثلثی باشند. تحت تعریف ۳ عملگر جمع، ضرب اسکالر و تفریق به شرح زیر تعریف می‌شوند (لی، ۲۰۰۵) :

الف: جمع $\tilde{\mu} \oplus \tilde{\nu} = (m_1 + m_2; a_1 + a_2, b_1 + b_2)_T$

ب: ضرب اسکالر

$$\lambda \otimes \tilde{\mu} = \begin{cases} (\lambda m_1; \lambda a_1, \lambda b_1)_T, & \lambda > 0, \\ (\lambda m_1; \lambda b_1, \lambda a_1)_T & \lambda < 0, \end{cases}$$

ج: تفریق $\tilde{\mu} \ominus \tilde{\nu} = \tilde{\mu} \oplus (-1 \otimes \tilde{\nu})$

تحت عملگرهای حسابی جمع و ضرب اسکالر اعداد فازی مثلثی یک فضای برداری خواهند بود.

تعریف ۴. (ژو و لی، ۲۰۰۱). فرض کنید $\tilde{\nu}$ و $\tilde{u}$ دو عدد فازی و $\tilde{\nu}_\alpha = [\nu_l(\alpha), \nu_r(\alpha)]$ و $\tilde{u}_\alpha = [u_l(\alpha), u_r(\alpha)]$ به ترتیب α ـبرش‌های $\tilde{\nu}$ و $\tilde{u}$ باشند. فاصله بین این دو عدد فازی بر پایه تابع وزنی $f(\alpha)$ به‌صورت

$$d(\tilde{u}, \tilde{\nu}) = \Big[\int_0^1 f(\alpha) d^2(\tilde{u}_\alpha, \tilde{\nu}_\alpha) d\alpha \Big]^{\frac{1}{2}},$$

تعریف می‌شود، که در آن $d^2(\tilde{u}_\alpha, \tilde{\nu}_\alpha) = [u_\ell(\alpha) - \nu_\ell(\alpha)]^2 + [u_r(\alpha) - \nu_r(\alpha)]^2$ و $f(\alpha)$ تابعی صعودی روی بازه $[0, 1]$ است، به قسمی که $\int_0^1 f(\alpha) d\alpha = \frac{1}{2}$ و $f(0) = 0$. در این حالت نرم عدد فازی $\tilde{u}$ به صورت $||\tilde{u}|| = d(\tilde{u}, 0)$ تعریف می‌گردد.

فرض کنید (Ω, F, P) یک فضای احتمال باشد. پوری و رالسکو (۱۹۸۶) تعریفی از متغیرهای تصادفی فازی را ارائه کردند که مبنای بیشتر مطالعات بعدی است.

تعریف ۵. تابع $\tilde{X}^* : \Omega \to F(\mathbb{R})$ را متغیر تصادفی فازی نامند، اگر اندازه‌پذیر (گات ، ۱۹۸۶) باشد. متغیر تصادفی فازی $\tilde{X}^*$ را کراندار انتگرال‌پذیر گویند اگر $E||\tilde{X}^*|| < \infty$.

لم ۱. (جو و کیم، ۲۰۰۱) فرض کنید $\tilde{X}^* : \Omega \to F(\mathbb{R})$ و $\tilde{X}^*(\omega) = \{(X_{\ell a}(\omega), (X_{ra}(\omega)) | 0 \le \alpha \le 1\}$ تحققی از $\tilde{X}^*$ باشد. $\tilde{X}^*$ یک متغیر تصادفی فازی است اگر و فقط اگر برای هر $0 \le \alpha \le 1$، $X_{\ell a}$ کران پایین آلفا برش $\tilde{X}^*$ و X_{ra} کران بالای آلفا برش $\tilde{X}^*$ متغیرهای تصادفی معمولی باشند.

متغیرهای تصادفی فازی مثلثی را می‌توان با سه متغیر تصادفی و به شکل $\tilde{X}^* = (X^L, X, X^U)$، $P(X^L < X < X^U) = 1$، توصیف کرد (حسامیان و اکبری ، ۲۰۲۲).

تعریف ۶. امید ریاضی متغیر تصادفی فازی کراندار انتگرال‌پذیر $\tilde{X}^*$ یک عدد فازی است که با

$$E(\tilde{X}^*) = \Big\{ \Big[\int X_{\ell a} dp, \int X_{ra} dp \Big] | 0 \le \alpha \le 1) \Big\},$$

تعریف می‌شود. در این صورت $E(\tilde{X}^*)(x) = \sup\left\{\alpha \in [0,1] : x \in E(\tilde{X}^*_\alpha)\right\}$.
فرض کنید $\mathbf{X}$بردار پیشگو و Y متغیر پاسخ نامنفی دو یا چند حالتی باشند. در بسیاری موارد ارتباط بین متغیر پاسخ با بردار پیشگو، $\mathbf{X}$ ، مدل‌بندی می‌شود. برای اینکه

$$Y = \boldsymbol{\beta}'\mathbf{X} \oplus \tilde{\epsilon}^*, \quad \boldsymbol{\beta} = (\beta_0, \ldots, \beta_m), \quad \mathbf{X} = (1, X_1, \ldots, X_m), \tag{۱}$$

که در آن $\boldsymbol{\beta}$ بردار پارامترهای مجهول و $\tilde{\epsilon}^*$ متغیر تصادفی فازی (مثلثی) است، یک الگوی مناسب برای مدل‌بندی ارتباط بین متغیر پاسخ و بردار پیشگو باشد، لازم است پارامترهای مدل برآورد شوند.

لم ۲. تحت شرایط (۱) اگر $\tilde{\epsilon}^*$ متغیر تصادفی فازی مثلثی با مرکز۰ باشد، Y به‌شرط $\mathbf{X}$ نیز متغیر تصادفی فازی مثلثی است.

برهان: به ازای هر $\omega \in \Omega$ و بنا به اصل گسترش $Y|\mathbf{X}$ یک عدد فازی مثلثی است. با توجه به خاصیت خطی بودن توابع اندازه‌پذیر، $Y|\mathbf{X}$ اندازه‌پذیر و لذا یک متغیر تصادفی فازی مثلثی است.

در مدل رگرسیون لوژستیک مقادیر متغیرهای پیشگو می‌توانند هر مقدار حقیقی باشند اما مقادیر متغیر پاسخ برچسب‌های رده‌بندی نامنفی (متغیرهای تصادفی فازی نامنفی) هستند، لذا به‌جای پیشگویی متغیر تصادفی پاسخ، که ماهیتی ابهامی نیز دارد، امید ریاضی آن، یعنی

$$\tilde{\mu} = E(Y|\mathbf{X}) = \boldsymbol{\beta}'\mathbf{X} \oplus \tilde{\epsilon}, \tag{۲}$$

مدل‌بندی می‌شود، که در آن $\tilde{\epsilon} = E(\tilde{\epsilon}^*)$ یک عدد فازی است. در این صورت $\tilde{\mu}$ نیز یک عدد فازی است.

۳ نتایج اصلی

تابع خطی سمت راست (۱) یک تابع بیکران است. متغیرهای پیشگو می‌توانند هر مقدار حقیقی باشند، لذا برای به الگو درآوردن مقدار متناهی (سمت چپ) مناسب نخواهد بود. در ادامه از تعمیم نظریه رگرسیون لوژستیک برای مدل‌بندی ارتباط بین $\tilde{\mu}$ و $\mathbf{X}$ استفاده خواهد شد. فرض می‌شود که $\tilde{\epsilon}$ یک عدد فازی LR ، $\tilde{\epsilon} = (m^*; a, b)_{LR}$ ، $a, b > 0$ باشد.

گزاره ۱. اگر در مدل (۱)، $\tilde{\epsilon} = (m^*; a, b)_{LR}$ ، $a, b > 0$ باشد، آنگاه تحت شرط $\mathbf{X} = \mathbf{x}$ ، $\tilde{\mu} = (m =$ ، $m^* + \boldsymbol{\beta}'\mathbf{x}; a, b)_{LR}$ ، $a, b > 0$ است.

برهان: بنا به اصل گسترش، $\tilde{\mu}$ یک مجموعه فازی است. چون $\tilde{\mu}$ تابع یک‌به‌یک است،

$$\tilde{\mu}(y) = \sup_{e, y=\boldsymbol{\beta}'\mathbf{x}+e} \tilde{\epsilon}(y - \boldsymbol{\beta}'\mathbf{x}) = \tilde{\epsilon}(e). \qquad (۳)$$

و لذا از ویژگی جمع اعداد مثلثی، نتیجه موردنظر حاصل می‌شود.
بدیهی است که در معادله (۲) تکیه‌گاه طرف چپ بازه صفر و یک، یا در حالت کلی یک بازه مثبت و متناهی، و طرف راست $\mathbb{R}$ است. لذا مدل‌بندی برای تبدیل (۴) بر نمونه تصادفی به حجم n اعمال می‌شود:

$$\tilde{w}_i = \ln(\frac{\tilde{\mu}_i}{۱ - \tilde{\mu}_i}) = \boldsymbol{\beta}'\mathbf{x}_i \oplus \tilde{\epsilon}_i. \qquad i = ۱, \ldots, n. \qquad (۴)$$

لم ۳. تحت شرایط مدل (۲) اگر $\tilde{\epsilon}$ یک عدد فازی LR ، $\tilde{\epsilon} = (m^*; a, b)_{LR}$ با تکیه‌گاه $[۰, ۱]$ باشد، آنگاه $\tilde{w} = \ln(\frac{\tilde{\mu}}{۱-\tilde{\mu}}) = \boldsymbol{\beta}'\mathbf{x} \oplus \tilde{\epsilon}$, یک عدد فازی با تابع عضویت زیر خواهد بود.

$$\tilde{w}(y) = \begin{cases} L(\frac{m - \frac{۱}{۱+e^{-y}}}{a}), & \frac{۱}{۱+e^{-y}} \leq m, \\ R(\frac{\frac{۱}{۱+e^{-y}} - m}{b}), & \frac{۱}{۱+e^{-y}} \geq m, \end{cases}$$

برهان: بنا به اصل گسترش

$$\tilde{w}(y) = \sup_{x \in [۰,۱], y = \ln \frac{x}{۱-x}} \tilde{\mu}(x) = \tilde{\mu}(\frac{۱}{۱ + e^{-y}}),$$

لذا بنا به گزاره ۱ حکم ثابت است.

فرع ۱. اگر $\tilde{\epsilon} = E(\tilde{\epsilon}^*) = (۰; \frac{۱}{۲})_T$ آنگاه $\tilde{\mu} = (m = \boldsymbol{\beta}'\mathbf{x}; \frac{۱}{۲})_T$،

$$\tilde{\mu}(x) = \begin{cases} ۱ - ۲|m - x|, & m - ۰/۵ \leq x \leq m, \\ ۱ - ۲|x - m|, & m < x < m + ۰/۵, \end{cases}$$

و $\tilde{\mu}_\alpha = [m + ۰/۵(\alpha - ۱), m + ۰/۵(۱ - \alpha)]$. اگر $\tilde{\epsilon} = (m^*; a, b)_T$ و $a, b > ۰$ آنگاه

$$\tilde{\mu}(x) = \begin{cases} ۱ - \frac{|m-x|}{a}, & m - a \leq x \leq m, \\ ۱ - \frac{|x-m|}{b}, & m < x \leq m + b. \end{cases}$$

فرع ۲. اگر $\tilde{\epsilon} = E(\tilde{\epsilon}^*) = (۰; \frac{۱}{۲})_T$، آنگاه آلفا برش‌ها $\tilde{w}$ به صورت

$$\tilde{w}_\alpha = \{y : \tilde{w}(y) \geq \alpha\} = \{y : \tilde{\mu}(\frac{۱}{۱ + e^{-y}}) \geq \alpha\}.$$

نتیجه می‌شود. با فرض آنکه $\tilde{\mu}$ عددی فازی با تکیه‌گاه مقادیر نامنفی است، $m + ۰/۵(\alpha - ۱) \geq ۰$ و

$$\begin{aligned}\tilde{w}_\alpha &= \{y : m + ۰/۵(\alpha - ۱) \leq \frac{۱}{۱ + e^{-y}} \leq m + ۰/۵(۱ - \alpha)\} \\ &= \{y : \ln \frac{m + ۰/۵(\alpha - ۱)}{۱ - m - ۰/۵(\alpha - ۱)} \leq y \leq \ln \frac{m + ۰/۵(۱ - \alpha)}{۱ - m - ۰/۵(۱ - \alpha)}\},\end{aligned}$$

یعنی

$$\tilde{w}_\alpha = [\ln \frac{(\alpha - ۱)۰/۵ + m}{۱ - ((\alpha - ۱)۰/۵ + m)}, \ln \frac{(۱ - \alpha)۰/۵ + m}{۱ - ((۱ - \alpha)۰/۵ + m)}] = [A_\ell, A_r]. \quad (۵)$$

در حالت کلی‌تر تحت شرایط لم ۳، $\tilde{\epsilon} = E(\tilde{\epsilon}^*) = (m^*; a, b)_T$، آنگاه الفا برش‌های $\tilde{w}$ نتیجه می‌شود

$$\begin{aligned}\tilde{w}_\alpha &= \{y : \tilde{w}(y) \geq \alpha\} = \{y : \tilde{\mu}(\frac{۱}{۱ + e^{-y}}) \geq \alpha\} \\ &= \{y : \ln \frac{m + b(\alpha - ۱)}{۱ - m - b(\alpha - ۱)} \leq y \leq \ln \frac{m + a(۱ - \alpha)}{۱ - m - a(۱ - \alpha)}\},\end{aligned}$$

فرع ۳. تحت شرایط مدل (۱) و با این فرض که $\tilde{\epsilon}^*$ متغیر تصادفی فازی مثلثی با مرکز ۰ باشد، مقدار مشاهده‌شده y به‌شرط $\mathbf{x}$ مقدار مشاهده‌شده از یک متغیر تصادفی فازی مثلثی به مرکز $\boldsymbol{\beta}'\mathbf{x}$ و پهنای متناظر با مقدار مشاهده‌شده از متغیرهای تصادفی X^L و X^U ، مثلاً h_ℓ و h_r ، است. یعنی در این حالت

$$\tilde{w}_\alpha = [\ln \frac{(\alpha - ۱)h_\ell + m}{۱ - ((\alpha - ۱)h_\ell + m)}, \ln \frac{(۱ - \alpha)h_r + m}{۱ - ((۱ - \alpha)h_r + m)}] = [A_\ell, A_r]. \quad (۶)$$

تحت شرایط مدل (۱) و با این فرض که $\tilde{\epsilon}^*$ متغیر تصادفی فازی مثلثی با مرکز e باشد، مقدار مشاهده‌شده y به‌شرط $\mathbf{x}$ مقدار مشاهده‌شده از یک متغیر تصادفی فازی مثلثی به مرکز $\boldsymbol{\beta}'\mathbf{x} + e$ است.

قضیه ۱. فرض کنید $(\mathbf{x}_۱, y_۱)$ ، ... ، $(\mathbf{x}_n, y_n)$ ، n مشاهده مستقل از مدل (۱) و $\tilde{\epsilon}_i^*$ ها متغیرهای تصادفی فازی مثلثی $\tilde{\epsilon}_i^* = (e_i; X_i^L, X_i^U)_T$ باشند، که در آن e_i مرکز خطا (و برابر صفر) و پهناهای X_i^L و X_i^U متغیرهای

تصادفی دارای توزیع یکنواخت در بازه $(۰,۱)$ است. در این صورت

$$\boldsymbol{\beta} = ۲(\sum_{i=۱}^{n} \mathbf{x}_i \mathbf{x}_i')^{-۱} \sum_{i=۱}^{n} z_i \mathbf{x}_i, \tag{۷}$$

برآورد کمترین توانهای دوم پارامتر $\boldsymbol{\beta}$ در مدل (۴) و نسبت به متر (۱) به ازای $f(\alpha) = \alpha$ است، که در آن m_i مرکز $\tilde{\mu}_i$ است و

$$\begin{aligned} z_i &= \int_{۰}^{۱} \alpha(\ln \frac{h_{i\ell}(\alpha - ۱) + m_i}{۱ - h_{i\ell}(\alpha - ۱) - m_i}) d\alpha + \int_{۰}^{۱} \alpha(\ln \frac{h_{ir}(۱ - \alpha) + m_i}{۱ - h_{ir}(۱ - \alpha) - m_i}) d\alpha \\ &= z_{i۱} + z_{i۲}. \end{aligned} \tag{۸}$$

برهان: تحت شرایط قضیه عدد فازی $E(\tilde{\epsilon}^*)$ به‌صورت $E(\tilde{\epsilon}^*) = (۰; \frac{۱}{۲}, \frac{۱}{۲})_T$ با آلفا برش‌های $(E(\tilde{\epsilon}^*))_\alpha =$ $[\frac{\alpha - ۱}{۲}, \frac{۱ - \alpha}{۲}]$. خواهد بود. کران‌های آلفا برش مقادیر مشاهده‌شده متغیر تصادفی فازی $\tilde{\epsilon}_i^* = (e_i; h_{i\ell}, h_{ir})_T$ به ترتیب $\tilde{\epsilon}_{\ell\alpha} = (-۱ + \alpha)h_{i\ell} + e_i$ و $\tilde{\epsilon}_{r\alpha} = (۱ - \alpha)(۱ - h_{ir}) + e_i$ است. در این صورت $\tilde{\mu}_i = (m =$ $\boldsymbol{\beta}'\mathbf{x}_i + e_i; \frac{۱}{۲}, \frac{۱}{۲})_T$. تحت شرایط مدل (۴)، اگر $\mathbf{X}_i = \mathbf{x}_i$ آنگاه

$$\begin{aligned} \tilde{\hat{w}}_i &= \boldsymbol{\beta}'\mathbf{x}_i + E(\tilde{\epsilon}^*) \\ \tilde{\hat{w}}_{i_\alpha} &= [\boldsymbol{\beta}'\mathbf{x}_i + e_i + \frac{(\alpha - ۱)}{۲}, \boldsymbol{\beta}'\mathbf{x}_i + e_i + \frac{(۱ - \alpha)}{۲}] = [B_{i\ell\alpha}, B_{ir\alpha}]. \end{aligned} \tag{۹}$$

پارامترهای مدل به روش کمترین مجموع توان دوم خطاها، به‌گونه‌ای برآورد می‌شود که فاصله بین بازه مشاهده‌شده $\tilde{w}_{i\alpha} = [A_{i\ell\alpha}, A_{ir\alpha}]$ و بازه برآورد شده $\tilde{\hat{w}}_{i\alpha} = [B_{i\ell\alpha}, B_{ir\alpha}]$، (که به‌اختصار از ذکر اندیس α خودداری

می‌گردد) مینیم شود. در این مرحله از فاصله (۱) استفاده می‌شود. در این صورت

$$\begin{aligned}
SSE &= \sum_{i=1}^{n} \int_0^1 \alpha d^2(\tilde{w}_{i\alpha}, \tilde{\hat{w}}_{i\alpha}) d\alpha \\
&= \sum_{i=1}^{n} \int_0^1 \alpha \Big([A_{i\ell} - B_{i\ell}]^2 + [A_{ir} - B_{ir}]^2 \Big) d\alpha \\
&= \sum_{i=1}^{n} \Big[\int_0^1 \alpha([A_{i\ell}^2 + A_{ir}^2] d\alpha + \int_0^1 \alpha(2(\boldsymbol{\beta}'\mathbf{x}_i)^2 + \frac{\alpha^2 - 2\alpha + 1}{2}) d\alpha \\
&- 2[\overbrace{k_{i1} + k_{i2}}^{k_i} + (\overbrace{z_{i1} + z_{i2}}^{z_i})(\boldsymbol{\beta}'\mathbf{x}_i)] \Big] \\
&= \sum_{i=1}^{n} \Big[\int_0^1 \alpha[A_{i\ell}^2 + A_{ir}^2] d\alpha + ((\boldsymbol{\beta}'\mathbf{x}_i)^2 + \frac{1}{24}) - 2[k_i + z_i(\boldsymbol{\beta}'\mathbf{x}_i)] \Big].
\end{aligned}$$

اکنون با مشتق‌گیری از SSE نسبت به $\boldsymbol{\beta}$ و برابر صفر قرار دادن آن رابطه $\boldsymbol{\beta} = 2(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i')^{-1} \sum_{i=1}^{n} z_i\mathbf{x}_i$ حاصل می‌شود. در حالت یک متغیره برآورد ضرایب به صورت زیر خواهند بود.

$$\hat{\boldsymbol{\beta}}_0 = \bar{z} - \hat{\beta}_1\bar{x}, \qquad \hat{\boldsymbol{\beta}}_1 = \frac{\sum_{i=1}^{n} z_i x_i - n\bar{x}\bar{z}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{۱۰}$$

تذکر ۱. قضیه ۱ را می‌توان به هر تابع چندجمله‌ای f و در حالت کلی‌تر هر تابع f که برای آن محاسبه z_iها امکان‌پذیر باشد تعمیم داد.

فرع ۴. اگر در مفروضات قضیه ۱ فرض پهناهای H_i متغیرهای تصادفی دارای توزیع یکنواخت در بازه $(0, 1)$ به فرض $H_i \sim U(\theta_1, \theta_2)$ تغییر یابد، با تکرار اثبات می‌توان نشان داد نتایج قضیه همچنان معتبر است. کافی است کران‌های انتگرال‌ها از $(0, 1)$ به (θ_1, θ_2) تغییر یابند.

۴ مثال‌های کاربردی

در این بخش داده‌ها در مورد بیماران قلبی است و متغیر زبانی به‌صورت دوحالتی (خیلی کم و خیلی زیاد)، متوسط و بیشتر از آن خیلی زیاد و کمتر از متوسط، خیلی کم در نظر گرفته می‌شود.

مثال ۱. بیماری قلبی: داده‌های بیماری قلبی، حاوی ۷۶ متغیر و ۳۰۳ مشاهده است، اما به دلایل مختلف تنها ۱۴ متغیر و در پایگاه داده‌ای کلیولند منتشرشده‌اند. این مجموعه داده از پایگاه داده‌ای UCI دریافت شده است. نام و

شماره بیمه بیماران از مجموعه داده‌ها حذف و مقادیر ساختگی جایگزین آن شده است. از مجموعه متغیرهای سن، جنسیت، نوع درد، فشارخون در حال استراحت،.... و وجود بیماری قلبی متغیر هدف یا پاسخ پنج متغیر پیشگو سن، فشارخون در حال استراحت، کلسترل سرم، ماکسیمم ضربان قلب، افسردگی و وجود بیماری قلبی مورد بررسی قرار می‌گیرند. اطلاعات مربوط به ۵ مشاهده نمونه در جدول ۱ ارائه شده است. متغیر پاسخ از ۰ (نداشتن بیماری قلبی) تا ۴ (داشتن بیماری قلبی شدید) ارزش‌گذاری شده است. برای سادگی محاسبات متغیر هدف دوتایی در نظر گرفته می‌شود. در این متغیر ۱ نشان‌دهنده (بیماری قلبی شدید) و ۰ (نداشتن بیماری قلبی شدید) است. در ۱۳

جدول ۱. نمایش ۵ مشاهده اول

Heart	Maximum	Cholestoral	Pressure	Blood_A	Type	Pain	Chest	Sex	Age
۰	۶	۱۵۰	۲	۱	۲۳۳	۱۴۵	۱	۱	۶۳
۲	۳	۱۰۸	۲	۰	۲۸۶	۱۶۰	۴	۱	۶۷
۱	۷	۱۲۹	۲	۰	۲۲۹	۱۲۰	۴	۱	۶۷
۰	۳	۱۸۷	۰	۰	۲۵۰	۱۳۰	۳	۱	۳۷

تا از مشاهدات مقدار متغیر هدف ۱ (گروه ۱) و در مابقی مشاهدات (۲۹۰ مشاهده) صفر (گروه ۲) است. یعنی داده‌ها خیلی نامتعادل هستند. برای ایجاد تعادل در مجموعه داده‌های آموزشی از روش نمونه‌گیری متناسب استفاده شده است. قاعده مشخصی برای نسبت نمونه‌های منتخب از رده‌های مختلف وجود ندارد اما معمولاً تلاش می‌شود که نسبت نمونه از هر رده (در حالت دو رده‌ای) تقریباً و حداقل ۲۰ درصد باشد. بنابراین از گروه اول ۱۲ نمونه و از گروه دوم ۵۸ نمونه انتخاب شده است. در این صورت نسبت نمونه در گروه اول برابر $P_1 = ۰/۹$ و در گروه دوم برابر $P_2 = ۰/۲$ است. در مجموعه داده‌های نمونه منتخب یک داده گم‌شده (مقدار متغیر الکتروکاردیوگرافی متغیر ۵۶) وجود دارد که با توجه به فراوانی مقادیر متغیر مقدار ۰ را به‌عنوان برآورد مقدار گم‌شده در نظر گرفته شده است. بر اساس اطلاعات هر یک از ۷۰ نمونه، از پزشک خواسته‌شده است تا امکان بیمار بودن آن‌ها را با توجه به علائم و وضعیت عمومی آن‌ها به‌صورت یکی از واژه‌های زبانی "بیمار بودن (۱) و بیمار نبودن (۰) بیان کند. پذیرفته می‌شود که نظر پزشک ماهیت تصادفی نیز دارد به‌گونه‌ای که در درجه عضویت واژه‌های زبانی تأثیرگذار است. معادله (۴) را می‌توان به‌صورت

$$\tilde{\mu}_i = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_i \oplus \tilde{\epsilon}_i))} = \frac{\exp(\beta_0 + \beta_1 X_i \oplus \tilde{\epsilon}_i)}{1 + \exp(\beta_0 + \beta_1 X_i \oplus \tilde{\epsilon}_i)}, \quad i = 1, \ldots, n.$$

نیز نوشت. شکل تعمیم‌یافته مانسکی و مک فادن (۱۹۸۱) مشخص‌سازی معادله (۴) را می‌توان به‌صورت

$$\tilde{\mu}_i = \frac{P_1}{P_2 + P_1 \exp(-(\beta_0 + \beta_1 X_i \oplus \tilde{\epsilon}_i))}, \qquad i = 1, \ldots, n.$$

بیان کرد. اگر $\gamma = \frac{P_1}{P_2}$ و $\gamma_1 = \ln(\frac{P_1}{P_2})$ آنگاه

$$\tilde{\mu}_i = \frac{\exp(\gamma_1 + \beta_0 + \beta_1 X_i \oplus \tilde{\epsilon}_i)}{1 + \exp(\gamma_1 + \beta_0 + \beta_1 X_i \oplus \tilde{\epsilon}_i)}, \qquad i = 1, \ldots, n.$$

اگر $\tilde{\epsilon} = E(\tilde{\epsilon}^*) = (0; \frac{1}{2}, \frac{1}{2})$، آنگاه از تعمیم نتیجه (۲) حاصل می‌شود:

$$\tilde{\mu}_\alpha = \{y : \tilde{\mu}(y) \geq \alpha\} = \{y : \tilde{\epsilon}_1(\ln \frac{y}{1-y}) \geq \alpha\}$$
$$= \{y : \frac{1}{\exp(-(\frac{1}{2}(\alpha - 1) + \gamma_1 + \boldsymbol{\beta}'\mathbf{x})) + 1} \leq y \leq \frac{1}{\exp(-(\frac{1}{2}(1 - \alpha) + \gamma_1 + \boldsymbol{\beta}'\mathbf{x})) + 1}\}.$$

فرض می‌شود که میزان بیماری قلبی اعداد فازی مثلثی به صورت

نداشتن بیماری قلبی شدید:

متغیر تصادفی فازی حدوداً ۰٫۳ تابع عضویت: $(0.3; u(0, 0.3); u(0.3, 0.6))_T$

بیماری قلبی شدید: متغیر تصادفی فازی حدوداً ۰٫۸ با تابع عضویت $(0.8; u(0.6, 0.8); u(0.8, 1))_T$ است. در این صورت بردارهای m ، x_ℓ و x_u از خواص اعداد فازی مثلثی به دست می‌آیند که به ترتیب عبارت‌اند از

$$m = \{0.3, 0.3, 0.3, \ldots, 0.8, 0.8\},$$
$$x_\ell = \{0.04, 0.23, 0.05, \ldots, 0.71, 0.67\},$$
$$x_r = \{0.60, 0.48, 0.40, \ldots, 0.93, 0.86\},$$

از تفریق بردارهای فوق، پهناهای چپ و راست حاصل می‌شوند.

$$h_\ell = m - x_\ell = 0.26, 0.07, 0.25, \ldots, 0.09, 0.13$$
$$h_r = x_r - m = 0.30, 0.18, 0.10, \ldots, 0.13, 0.06.$$

در پرونده داده‌ها نخست نمونه‌های فاقد بیماری قلبی شدید و سپس داده‌های دارای بیماری قلبی شدید آورده شده است. پس از برآورد پارامترها، مدل به‌صورت

$$\tilde{w}_i = -0.495 + 0.023X_1 + 0.009X_2 0.001X_3 - 0.024X_4 + 0.235X_5 \oplus \tilde{\epsilon}.$$

جدول ۲. جدول مربوط به پهناها

h_r	h_ℓ	x_r	x_ℓ	m
۰٫۳۰	۰٫۲۶	۰٫۰۶	۰٫۰۴	۰٫۳۰
۰٫۱۸	۰٫۰۷	۰٫۴۸	۰٫۲۳	۰٫۳۰
۰٫۱۰	۰٫۲۵	۰٫۴۰	۰٫۰۵	۰٫۳۰
۰٫۱۷	۰٫۰۴	۰٫۴۷	۰٫۲۶	۰٫۳۰
۰٫۰۱	۰٫۱۷	۰٫۳۱	۰٫۱۳	۰٫۳۰

حاصل شده است. برای اولین داده آزمونی، مقدار پیش‌بینی به صورت

$$\tilde{w}_i = -\ ۰٫۴۹۴۹ + ۰٫۰۲۳۵ \times ۶۹ + ۰٫۰۰۹۳ \times ۱۴۵ + ۰٫۰۰۱۰ \times ۲۳۳$$
$$- ۰٫۰۲۳۶ \times ۱۵۰ + ۰٫۲۳۵۰ \times ۲٫۳ \oplus \tilde{\epsilon} = ۰٫۶۸۲ \oplus \tilde{\epsilon},$$

و آلفا برش‌ها به صورت

$$\tilde{\mu}_\alpha = \Big[\frac{۱}{۱ + \exp(-((\alpha - ۱)۰٫۵ - ۰٫۶۸۲)}, \frac{۱}{۱ + \exp(-((۱ - \alpha)۰٫۵ - ۰٫۶۸۲)}\Big][\tilde{A}_L^\alpha, \tilde{A}_U^\alpha],$$

خواهند بود. با استفاده از معیار غیر فازی‌ساز $E_{\tilde{A}} = \int_۰^۱ (۱ - ۲^{-\alpha})(\tilde{A}_L^\alpha + \tilde{A}_U^\alpha) d\alpha$، خروجی به عدد حقیقی تبدیل می‌شود. در حالت فوق مقدار خروجی ۰٫۴۵۵ خواهد بود. اگر هنگامی که مقدار خروجی کمتر از۰٫۵ باشد، رده صفر در نظر گرفته شود، الگوریتم فوق به درستی رده اولین مشاهده آزمونی را تعیین می‌کند. با استفاده از این الگوریتم رده هر یک از مشاهدات آزمونی، شامل ۲۳۳ مشاهده، تعیین شدند. نرخ خطای کل یا به طور ساده‌تر نرخ خطا که برابر است با مجموع منفی‌ها و مثبت‌های کاذب تقسیم بر تعداد کل مشاهدات برابر است با $\frac{۹+۱}{۲۳۳}$. به همین صورت نرخ منفی کاذب، تعداد منفی‌های کاذب تقسیم بر تعداد کل طبقه‌بندی‌های منفی، طبقه‌بندی صفر، برابر $\frac{۹}{۲۳۲}$ است. جدول ۳ نرخ خطای کل، نرخ منفی‌های کاذب و نرخ مثبت‌های کاذب را نشان می‌دهد. کد برنامه در آدرس https://github.com/maryammaleki68/Fuzzy-logit/edit/main/README.md قابل دسترس است.

جدول ۳. خلاصه نتایج برای داده آزمودنی

	تشخیص مدل		
واقعی	عدم ابتلا	ابتلا	جمع
عدم ابتلا	۲۲۳	۹	۲۳۲
ابتلا	۱	۰	۱

مثال ۲. شبیه سازی:در این مثال متغیرهای سن، فشارخون در حال استراحت، کلسترل سرم، ماکسیمم ضربان قلب و افسردگی، هر یک با استفاده از توزیع یکنواخت در بازه مقادیر متغیرهای پیشگو که در زیر آمده است، تولید شده و برآورد پارامترها محاسبه شده است. همچنین مقدار کمترین مربعات خطا را برای مقادیر β برآورد شده از داده‌های اصلی و داده‌های شبیه سازی شده محاسبه و برابر ۰٫۸۵۲۷ است. کدهای مربوط به برنامه شبیه‌سازی در بخش پیوست آمده است. متغیر سن دارای توزیع $U(۳۳, ۸۰)$، متغیر فشارخون در حال استراحت دارای توزیع $U(۱۰۰, ۲۰۰)$، متغیر کلسترل سرم دارای توزیع $U(۱۴۰, ۵۷۰)$، متغیر ماکسیمم ضربان قلب دارای توزیع $U(۱۱۰, ۲۰۰)$ و متغیر افسردگی دارای توزیع $U(۰, ۶٫۵)$ است.

در ادامه روابطی بازگشتی برای برآورد پارامترها به دست آورده شده است. این روابط بازگشتی می‌توانند ابزار مهمی در یادگیری ماشین و برای داده‌های بزرگ باشند. در ادامه از نماد $\bullet$ به‌جای $n+۱$ استفاده می‌شود. مثلاً $\bar{z}^{\bullet}$ نشان‌دهنده میانگین z_i به ازای حجم نمونه برابر $n+۱$ است. در این صورت در رابطه $\hat{\beta}_{۰}^{\bullet} = \bar{z}^{\bullet} - \hat{\beta}_{۱}^{\bullet}\bar{x}^{\bullet}$ خواهیم داشت:

$$\bar{z}^{\bullet} = \frac{n}{n+۱}\bar{z} + \frac{z^{\bullet}}{n+۱}, \quad \bar{x}^{\bullet} = \frac{n}{n+۱}\bar{x} + \frac{x^{\bullet}}{n+۱}.$$

در نتیجه

$$\hat{\beta}_{۰}^{\bullet} = \frac{n}{n+۱}(\bar{z} - \hat{\beta}_{۱}^{\bullet}\bar{x}) + \frac{z^{\bullet}}{(n+۱)} - \hat{\beta}_{۱}^{\bullet}\frac{x^{\bullet}}{(n+۱)}.$$

که در آن $z^{\bullet}$ ، $x^{\bullet}$ مربوط به داده جدید است. همچنین برآورد $\hat{\beta}_{۱}^{\bullet}$ نیز به‌صورت

$$\begin{aligned}\hat{\beta}_{۱}^{\bullet} &= \frac{\sum_{i=۱}^{n+۱} z_i x_i - (n+۱)\bar{x}^{\bullet}\bar{z}^{\bullet}}{\sum_{i=۱}^{n+۱} x_i^{۲} - (n+۱)\bar{x}^{\bullet ۲})} \\ &= \frac{\sum_{i=۱}^{n} z_i x_i - (\frac{n}{(n+۱)}n\bar{x}\bar{z} + \frac{n}{(n+۱)}\bar{x}z^{\bullet} + \frac{n}{n+۱}x^{\bullet}\bar{z} + \frac{z^{\bullet}x^{\bullet}}{n+۱}) + x^{\bullet}z^{\bullet}}{\sum_{i=۱}^{n} x_i^{۲} - (n+۱)(\frac{n}{(n+۱)^{۲}}n\bar{x}^{۲} + ۲\frac{n}{(n+۱)^{۲}}\bar{x}x^{\bullet} + \frac{x^{\bullet ۲}}{(n+۱)^{۲}}) + x^{\bullet ۲}},\end{aligned}$$

است. برای مقادیر بزرگ n ، $\frac{n}{(n+1)} \cong 1$ و $\frac{1}{(n+1)} \cong 0$ ، لذا

$$\hat{\beta}_1^\bullet \cong \frac{\overbrace{(\sum_{i=1}^{n} z_i x_i - n\bar{x}\bar{z})}^{k} - \overbrace{(\bar{x}z^\bullet + x^\bullet \bar{z} - x^\bullet z^\bullet)}^{\theta}}{\underbrace{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}_{m} \underbrace{-2\bar{x}x^\bullet + x^{\bullet 2}}_{\Omega}} \cong \hat{\beta}_1 - \frac{k\Omega + m\theta}{m(m+\Omega)}.$$

بنابراین

$$\hat{\beta}_1^\bullet \cong \hat{\beta}_1 - \frac{(\sum_{i=1}^n z_i x_i - n\bar{x}\bar{z})(-2\bar{x}x^\bullet + x^{\bullet 2}) + (\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\bar{x}z^\bullet + x^\bullet \bar{z} - x^\bullet z^\bullet)}{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n x_i^2 - n\bar{x}^2 - 2\bar{x}x^\bullet + x^{\bullet 2})}.$$

به‌طور مشابه و با جایگذاری تقریب $\hat{\beta}_1^\bullet$ در (۱۱) برای پارامتر $\hat{\beta}_0^\bullet$ خواهیم داشت:

$$\begin{aligned}
\hat{\beta}_0^\bullet &= \frac{n}{(n+1)}(\bar{z} - \hat{\beta}_1^\bullet \bar{x}) + \frac{z^\bullet}{(n+1)} - \hat{\beta}_1^\bullet \frac{x^\bullet}{(n+1)} \\
&\cong \hat{\beta}_0 \\
&- \frac{(\sum_{i=1}^n z_i x_i - n\bar{x}\bar{z})(-2\bar{x}x^\bullet + x^{\bullet 2}) + (\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\bar{x}z^\bullet + x^\bullet \bar{z} - x^\bullet z^\bullet)}{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n x_i^2 - n\bar{x}^2 - 2\bar{x}x^\bullet + x^{\bullet 2})} \bar{x} \\
&+ \frac{z^\bullet}{(n+1)} - (\hat{\beta}_1 \\
&- \frac{(\sum_{i=1}^n z_i x_i - n\bar{x}\bar{z})(-2\bar{x}x^\bullet + x^{\bullet 2}) + (\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\bar{x}z^\bullet + x^\bullet \bar{z} - x^\bullet z^\bullet)}{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n x_i^2 - n\bar{x}^2 - 2\bar{x}x^\bullet + x^{\bullet 2})}) \\
&\frac{x^\bullet}{(n+1)}.
\end{aligned}$$

بحث و نتیجه‌گیری

رگرسیون لوژستیک یکی از روش‌های یادگیری بر پایه آمار و احتمال و ابزار مناسبی برای رده بندی مشاهدات است. در این روش و بر مبنای متغیرهای پیشگو مقدار متغیر وابسته که متغیری رسته‌ای یا اسمی است، برآورد می‌شود. تعیین متغیر وابسته معادل تعیین رده‌ای است که مشاهده به آن رده تعلق دارد. اما در مطالعات اجتماعی، روانشناسی، بالینی، ...این امکان وجود دارد که متغیرهای پیشگو یا وابسته یا پارامترهای مدل ماهیتی ابهامی داشته باشند. حالتی که متغیرهای وابسته، مستقل یا پارامترهای مدل فازی باشند، توسط مؤلفین مختلف و تحت شرایط

متنوعی موردبررسی قرار گرفته‌اند. در این مقاله حالتی بررسی شده است که متغیر پاسخ علاوه بر ماهیت فازی بودن ماهیت تصادفی نیز داشته باشد. در این مقاله فرض شده که متغیر پاسخ به دلیل وجود خطای تصادفی فازی، متغیر تصادفی فازی مثلثی باشد و پارامترهای مدل رگرسیون لوژستیک برآورد شدند. انتظار می‌رود که این مقاله چشم انداز گسترده‌ای را برای مطالعه و کاربرد رگرسیون لوژستیک هنگامی که برخی یا همه متغیرها تصادفی فازی باشند، ایجاد کند. اگر حجم نمونه بزرگ باشد، محاسبه برآوردها مشکل خواهد بود. در انتها روشی بازگشتی برای محاسبه پارامترها به دست آمده است. این روابط، کاربرد روش پیشنهادی برای داده‌های بزرگ را امکان‌پذیر می‌سازند. با یک مثال به تشریح نتایج به دست آمده پرداخته شده است. در این مثال داده‌ها با استفاده از نمونه گیری متناسب به دو دسته، داده‌های آزمایشی و داده‌های آزمونی تقسیم شده‌اند. بر مبنای داده‌های آزمایشی پارامترها برآورد شدند. با استفاده از پارامترهای برآورده شده رده اولین مشاهده از داده‌های آزمونی (و به درستی) تعیین گردید.

تقدیر و تشکر

مؤلفین از داوران، سردبیر و ویراستار محترم که پیشنهادات ارزشمند ایشان موجب بهبود مقاله گردید کمال تشکر و سپاسگزاری را دارند.

مراجع

طاهری، س. م. و ماشین چی، م. (۱۳۸۷)، مقدمه‌ای بر احتمال و آمار فازی، دانشگاه شهید باهنر کرمان.

مقیم بیگی، م. (۱۴۰۱)، رگرسیون لوژستیک چند جمله‌ای نیمه پارامتری برای رده‌بندی داده‌های شکل، مجله علوم آماری، ۱۶(۲)، ۴۴۹-۴۶۸.

Agresti, A. (2015), *Foundations of Linear and Generalized Linear Models*, Hoboken: Wiley.

Fauci, A., Braunwald, S. E., Kasper, D. L., Hauser, S. L., Longo, D. L., Jameson, J. L., and Loscalzo, J. (2008), Harrison's Principals of Internal Medicine, New York: Wiley. 2275- 2279.

Goetschel, R., and Voxman, W. (1986), Elementary Fuzzy Calculus, *Fuzzy Sets and Systems*, **18**, 31-43.

Gut, A. (2005), *Probability, A Graduated Course*, Springer Science+Business Media, Inc.

Hesamian, G. and Akbari, M. G. (2017), Semi-Parametric Partially Logistic Regression Model with Exact Inputs and Intuitionistic Fuzzy Output Applied, *Soft Computing*, **58**, 517-526.

Hesamian, G. and Akbari, M. G. (2022), A Fuzzy Empirical Quantile-Based Regression Model Based on Triangular Fuzzy Numbers, *Computational and Applied Mathematics*, **41**(6), 1-26

Joo, S. Y., and Kim, Y. K. (2001), Kolmogorov's Strong Law of Large Numbers for Fuzzy Random Variables, *Fuzzy Sets and Systems*, **120**(3), 499-503.

Lee, K. H. (2005), *First Course on Fuzzy Theory and Applications*, Berlin: Springer.

Mirzaei Yeganeh, S., and Taheri, S.M. (2010), Possibilistic Logistic Regression by Linear Programming Approach, *Proc of 7th Probability and Random Process*, Isfahan University of Technology, Isfahan, Iran, 139-143.

Mohd Dom, R. (2009), A Fuzzy Regression Model for The Prediction of Oral Cancer Susceptibility. Ph.D. Thesis, *Computer Science and Information Technology, University of Malaya*, Kualalumpur.

Manski, C., and McFadden, D. (1981), Alternative Estimators and Sample Designs for Discrete Choice Analysis. In: Manski, McFadden,Editors, *Structural Analysis of Discrete Data with Econometric Applications*, London: MIT Press.

Nagar, P., and Srivastava, S. (2008), Adaptive Fuzzy Regression Model for The Prediction of Dichotomous Response Variables Using Cancer Data: A Case Study, *Journal of Applied Mathematics Statistics and Informatics*, **4**, 183-191.

Namdari, M., Taheri, S.M., Abadi, A., Rezaei, M., and Kalantari, N. (2013), Possibilistic Logistic Regression for Fuzzy Categorical Response Data, *International Conference on Fuzzy Systems*, **8**, 1-6.

Namdari, M., Yoon, J. H., Abadi, A., Taheri, S. M., and Choi, S. H. (2015), Fuzzy Logistic Regression with Least Absolute Deviations Estimators, *Soft Computing*, **19**, 909-917.

Pourahmad, S., Ayatollahi, S. M. T., Taheri, S. M., and Agahi, Z. H. (2011.a), Fuzzy Logistic Regression Based on The Least Squares Approach with Application in Clinical Studies, *Computer Mathematics Applications*, **62**, 3353-3365.

Pourahmad, S., Ayatollahi, S. M., and Taheri, S. M. (2011.b), Fuzzy Logistic Regression, A New Possibilistic Regression and Its Application in Clinical Vague Status, *Iranian Journal of Fuzzy Systems*, **8**, 1-17.

Puri, M. L., and Ralescu, D. A. (1986), Fuzzy Random Variables, *Journal of Mathematical Analysis and Applications*, **114**, 409-422.

Salmani, F., Taheri, S. M., Yoon, J. H., Abadi, A., Alavi Majd, H., and Abbaszadeh, A. (2017), Logistic Regression for Fuzzy Covariates: Modeling, Inference and Applications, *International Journal of Fuzzy Systems*, **19**, 1635-1644.

Taheri, S. M., and Mirzaei Yeganeh, S. (2009), Logistic Regression with Non-Precise Response, *Proceedings of The 57th ISI Conference, Durban, South Africa*, 98-101.

Takemura, K. (2004), Fuzzy Logistic Regression Analysis for Fuzzy Input -Output Data, *International Conference on Soft Computing and Intelligent Systems and the 5th International Symposium on Advanced Intelligent Systems, Japan*, 1-6.

Tseng, F., and Lin, L. (2005), A Quadratic Interval Logit Model for Forecasting Bankruptcy, *Omega*, **33**, 85–91.

Yu, J. R., and Tseng, F. M. (2014), Fuzzy Piecewise Logistic Growth Model for Innovation Diffusion: A Case Study of The TV Industry, *International Journal of Fuzzy Systems*, **18**, 1-12.

Xu, R., and Li, C. (2001), Multidimensional Least-Squares Fitting with A Fuzzy Model, *Fuzzy Sets and Systems*, **119**(2), 215-223.