برآورد پارامترهای مدل رگرسیون لوژستیک به کمک ماکسیمم آنتروپی تعمیم‌یافته

مهسا مرکانی [۱] ، منیژه صانعی‌طبس [۲] ، حبیب نادری [۳] ، حامد احمدزاده [۴] ، جواد جمال‌زاده [۵]

تاریخ دریافت: ۱۴۰۰/۰۷/۲۵
تاریخ پذیرش: ۱۴۰۰/۱۲/۲۶

چکیده:
هنگام کار با یک مجموعه داده رگرسیونی ممکن است برخی شرایط برقرار نباشند و محدودیت‌هایی برای اجرای مدل رگرسیون به وجود آیند که ما را در استفاده از روش حداقل مربعات با مشکل مواجه می‌کند. روش ماکسیمم آنتروپی تعمیم‌یافته با زیربنای رگرسیونی قادر است پارامترهای مدل رگرسیونی را بدون در نظر گرفتن هیچ شرطی روی توزیع احتمال خطاها برآورد کند. پیش‌ازاین توانمندی این روش در حجم نمونه‌های کم بررسی و تأییدشده است. هنگامی‌که متغیر پاسخ یک متغیر کیفی است روش رگرسیون لوژستیک به کار می‌رود در این پژوهش ابتدا روش ماکسیمم آنتروپی تعمیم‌یافته را برای یک مدل رگرسیونی لوژستیک معرفی کردیم. یک نمونه تصادفی از مشتریان بانک جمع‌آوری‌شده و در این بررسی برای برآورد پارامترهای مدل از مدل رگرسیون لوژستیک دودویی و با استفاده از دو روش ماکسیمم آنتروپی تعمیم‌یافته و درستنمایی ماکسیمم تجزیه‌وتحلیل و کار آماری انجام گرفت و درنهایت دو روش ذکرشده را باهم مقایسه می‌کنیم. با استناد به مقدار آماره میانگین مربعات خطا برای پیش‌گویی تقاضای مشتری برای افتتاح حساب بلند مدّت که از رگرسیون لوژستیک با استفاده از دو روش ماکسیمم آنتروپی تعمیم‌یافته و درستنمایی ماکسیمم به‌دست‌آمده، مشخص شد که روش برآورد ماکسیمم آنتروپی تعمیم‌یافته دارای نتایج دقیق‌تری است.

واژه‌های کلیدی: رگرسیون لوژستیک، ماکسیمم آنتروپی تعمیم‌یافته، درستنمایی ماکسیمم، میانگین مربعات خطا

۱ مقدمه

وقتی یک مدل رگرسیونی می‌خواهیم به یک مجموعه داده برازش دهیم باید یکسری پذیره‌های زیربنایی شامل ۱) نرمال بودن باقیمانده‌ها ۲) صفر بودن میانگین باقیمانده‌ها ۳) همگنی واریانس باقیمانده‌ها ۴) استقلال باقیمانده‌ها برقرار باشند که عدم وجود هر یک از این شرایط ما را در به کار بردن روش رگرسیونی حداقل مربعات با مشکل مواجه می‌کند.

رگرسیون لوژستیک[۶] برای تحلیل روابط بین متغیر پاسخ و یک یا چند متغیر توضیحی به کار می‌رود، این نوع رگرسیون به‌عنوان جایگزینی برای روش طبقه‌بندی فیشر و تحلیل تشخیصی به‌کاررفته است. این نوع از رگرسیون رابطه بین چند متغیر توضیحی با یک متغیر پاسخ دودویی[۷] را مورد تجزیه‌وتحلیل قرار می‌دهد. در اینجا فقط دو حالت ممکن برای متغیر پاسخ وجود دارد، موفقیت یا عدم موفقیت، وقوع یک رخداد یا عدم وقوع آن و ... .

روش درستنمایی ماکسیمم[۸] (ML) محبوب بسیاری از محققین بوده و برای برآورد بسیاری از مسائل به کار می‌رود. نقطه شروع این تاریخچه را در کارهای پیرسون برای اندازه‌گیری نیکویی برازش یک مدل و روش گشتاور پیرسون باید جستجو کرد [۱۳] و[۱۴]. این منطق توسط فیشر با کارش بر روی روش ML ادامه پیدا کرد [۴] و [۵]. امّا وقتی پذیره‌های زیربنایی رگرسیونی برقرار نباشد باید به دنبال یک روش جایگزین برای برآورد باشیم. جاج و گلان پارامترهای مدل رگرسیونی

[۱] دانشجوی کارشناسی ارشد دانشگاه سیستان و بلوچستان
[۲] عضو هیئت‌علمی دانشگاه سیستان و بلوچستان
[۳] عضو هیئت‌علمی دانشگاه سیستان و بلوچستان (نویسنده مسئول: h.h.naderi@gmail.com)
[۴] عضو هیئت‌علمی دانشگاه سیستان و بلوچستان
[۵] عضو هیئت‌علمی دانشگاه سیستان و بلوچستان

[6] Logistic Regression
[7] Binary
[8] Maximum Likelihood

را به کمک روش ماکسیمم آنتروپی تعمیم‌یافته[9] (GME) در مسائل اقتصادی برآورد کردند ([۱۰]). سپس گلان و همکاران در سال ۱۹۹۶ روش GME را معرفی کردند و نشان دادند که برای استفاده از این روش نیازی به برقراری پذیره‌های یادشده نیست و پس‌ازآن محققان مختلف در مسائل مختلف از این روش استفاده کردند که از آن جمله می‌توان به موارد زیر اشاره کرد:

سیاولینو و کالکانی در پژوهشی که دارای حجم نمونه کم و هم‌خطی چندگانه بود، روش برآورد GME را استفاده کردند ([۳]). سریبوونچیتا و همکاران در مسائل اقتصادی از روش GME بهره بردند ([۱۷]). یادهیسا و همکاران در پژوهشی که دارای حجم نمونه کم بود، روش GME را استفاده کردند ([۱۸]). سانگ و ویت مدل اقتصادسنجی نوین را برای برآورد و پیش‌بینی تقاضای گردشگری ارائه دادند و به روندهای تحقیق مربوط به پیش‌بینی تقاضای گردشگری اشاره کردند ([۱۶]). هدف از انجام این پژوهش مقایسه برآورد روش GME و برآورد روش ML برای مدل رگرسیون لوژستیک است. با توجه به اهمیت و کاربرد روش GME، در این مطالعه تلاش گردید تا قابلیت این روش با روش ML با استفاده از آماره میانگین مربعات خطا[10] (MSE) مشخص شود. فرض کنید از مشتریان یک بانک در خصوص اینکه آیا مجدداً تصمیم به افتتاح حساب سپرده بلند مدّت در این بانک دارند یا خیر پرسیده شده است، برخی جوابشان مثبت و برخی جواب منفی داده‌اند و بنا به هر دلیلی ترجیح داده‌اند در این بانک حساب سپرده بلند مدّت نداشته باشند. پس متغیر پاسخ یک متغیر کیفی دوحالتی است. نمونه‌ای به حجم ۳۹۹ نفر به‌طور تصادفی شامل اطلاعاتی از قبیل سن، شغل، وضعیت تأهل، میزان تحصیلات و وام (منظور این است که آیا مشتری وام گرفته است یا خیر.) به‌عنوان متغیرهای توضیحی و تقاضای مشتری برای افتتاح حساب بلند مدّت به‌عنوان متغیر پاسخ گرفته شد. با توجه به اینکه متغیر پاسخ، یک متغیر دودویی است رگرسیون لوژستیک باید به کار گرفته شود و برای برآورد پارامترهای مدل از روش برآورد GME استفاده شد. همچنین با برآورد پارامترهای مدل با استفاده از روش ML و استناد به آماره MSE مشخص شد که روش GME از دقّت بالاتری برخوردار است. لذا در ادامه با توجه به اینکه مدل رگرسیون این پژوهش از نوع لوژستیک است روش برآورد مربوطه برای مدل لوژستیک شرح داده‌شده است.

۲ معرفی روش ماکسیمم آنتروپی تعمیم‌یافته در مدل رگرسیونی لوژستیک

روش GME برای اولین بار با چارچوب رگرسیونی مطرح و به کار گرفته شد. این روش ضمن دوباره پارامتری کردن مدل خطی مربوطه پارامترها و خطاهای مجهول را به فرم احتمالاتی که باید برآورد شوند در نظر می‌گیرد. بدین ترتیب توزیع احتمال هر یک از عناصر مجهول برآورد گردیده و مجموع آنتروپی‌های توزیع ضرایب و خطاهای متناظر ماکسیمم می‌شود. مدل رگرسیون خطی زیر با T مشاهده و K متغیر توضیحی در نظر بگیرید:

$$Y = X\beta + e, \qquad (۱)$$

که در آن Y یک بردار T بعدی از متغیر تصادفی مشاهده‌شده، X یک ماتریس $T \times K$ از متغیرهای توضیحی و $\beta = (\beta_۱, ..., \beta_K)$ یک بردار K بعدی از پارامترهای مجهول است که از روی داده‌ها برآورد می‌شوند و e نیز بردار T بعدی از خطاهای تصادفی غیرقابل مشاهده هستند. برای تعیین تکیه‌گاه خطا، گلان و همکاران از قانون ۳σ پولکیشم[11] استفاده کرده‌اند کران پایین را -۳σ و کران بالا را $+۳\sigma$ قرار داده که مقدار σ نیز انحراف معیار بردار Y است ([۱۵]).

بنابراین در ادامه به کمک تکیه‌گاه پیشین، برآوردهای GME ضرایب رگرسیونی و جملات خطا به‌صورت،

$$\beta_k^{GME} = \sum_{m=۱}^{M} \hat{p}_{km} z_{km}, \ k = ۱, ۲, ..., K. \qquad (۲)$$

$$e_t^{GME} = \sum_{j=۱}^{J} \hat{w}_{tj} v_{tj}, \ t = ۱, ۲, ..., I. \qquad (۳)$$

حاصل می‌شوند. ملاحظه می‌شود که جواب‌های β_k^{GME} و e_t^{GME} هر دو به ضرایب لاگرانژ λ_t وابسته بوده و هیچ فرم بسته‌ای برای آن‌ها وجود ندارد و باید مقدار آن‌ها به روش‌های عددی محاسبه گردد ([۱]). در مواردی که متغیر پاسخ کیفی است رگرسیون لوژستیک موردنیاز است. حال برای برآورد ضرایب رگرسیون لوژستیک چه باید کرد؟

در تحلیل رگرسیون لوژستیک، همیشه یک متغیر پاسخ و معمولاً مجموعه‌ای از متغیرهای توضیحی وجود دارند که ممکن است دودویی،

[9] Generalized Maximum Entropy

[10] Mean Square Error

[11] Pullkishm

کمی یا ترکیبی از آن‌ها باشند. به‌علاوه لازم نیست متغیرهای دودویی به‌طور واقعی دوتایی باشند. برخی از متغیرهای توضیحی در رگرسیون لوژستیک می‌توانند به‌عنوان متغیرهای کمکی مورداستفاده قرار گیرند تا پژوهشگران بتوانند با ثابت نگه‌داشتن یا کنترل آماری این متغیرها اثرات دیگر متغیرهای توضیحی را بهتر ارزیابی کنند.

با این‌که رگرسیون لوژستیک در مقایسه با رگرسیون خطی پیش‌فرض‌های کمتری دارد (به‌عنوان‌مثال پیش‌فرض‌های همگنی واریانس و نرمال بودن خطاها وجود ندارد)، رگرسیون لوژستیک نیازمند موارد زیر است:

۱. هم‌خطی چندگانه کامل وجود نداشته باشد.

۲. خطاهای خاص نباید وجود داشته باشد (یعنی همه متغیرهای پیش‌بین مرتبط وارد شوند و پیش‌بین‌های نامربوط کنار گذاشته شوند.)

۳. متغیرهای توضیحی باید در مقیاس فاصله‌ای یا سطح نسبی اندازه‌گیری شده باشند (هرچند که متغیرهای دودویی نیز می‌توانند مورداستفاده قرار گیرند.)

دو استفاده مهم از رگرسیون لوژستیک عبارت‌اند از:

۱- پیش‌بینی اعضای گروه: در رگرسیون لوژستیک احتمال موفقیت قبل از احتمال شکست محاسبه می‌شود، بنابراین نتایج به فرم نسبت شانس به‌دست‌آمده می‌آیند.

۲- شناسایی و اطلاع از روابط و شدت آن در بین متغیرها.

در این بخش روش GME برای مدل رگرسیون لوژستیک را بیان می‌کنیم. همچنان که قبلاً متذکر شدیم با توجه به اینکه وزن‌های مجهول پارامترها و تکیه‌گاه خطاهای مدل رگرسیونی از یکدیگر مستقل‌اند [۶]، [۷] و [۸] می‌توان با حل مسئله بهینه‌سازی ماکسیمم آنتروپی زیر، پارامترها و خطاهای مجهول را برآورد کرد. اطلاع محدودیت گشتاوری را در مسئله آنتروپی باید به کار ببندیم پس توابع غیرقابل مشاهده و مجهول p و e خواص تابع احتمال را باید داشته باشند. عناصر p که به فرم احتمالات هستند و امّا اعداد حقیقی مقدار e_{ij} روی بازه $[-۱,۱]$ تغییر می‌کند. چون این خطاها به فرم احتمال نیستند، یک ابزار مفهومی بازنویسی پارامتری که توسط جاج (۱۹۹۱) و جاج و گلان (۱۹۹۳) برای تبدیل e_{ij} ها به مقادیر احتمالی که روی بازه $(۰,۱)$ تغییر کنند را به کار می‌گیریم پس‌روی بازه $[-۱,۱]$ متغیر تصادفی گسسته کران‌دار با مقادیر ممکن اعداد حقیقی $\nu_{ij}=(\nu_{ij۱},\nu_{ij۲},...,\nu_{ijH})$ با بعد $H\geq ۲$ تعریف می‌کنیم. با وزن‌های متناظر مجهول w_{ij} که $w_{ij}=(w_{ij۱},w_{ij۲},...,w_{ijH})$ که دارای خواص تابع احتمال هستند و باید $\sum_{h=۱}^{H}w_{ijh}=۱$ و بنابراین $e_{ij}=\sum_{h=۱}^{H}w_{ijh}\nu_h$ مثلاً اگر قرار دهیم $H=۳$ آنگاه $\nu=(-۱,۰,۱)$ پس نقاط $w_۱$ و $w_۲$ و $w_۳$ وجود دارند به‌طوری‌که مقادیری مثبت هستند و نیز هر e_{ij} را می‌نویسیم،

$$e_{ij}=w_۱(-۱)+w_۲(۰)+w_۳(۱)=w_۱(-۱)+w_۳(۱)$$

نقش H، تعیین مقدار H و مثال‌های تئوری و بحرانی که در آن‌ها H چگونه انتخاب‌شده است در مقاله گلان و همکاران (۱۹۹۶) آورده شده است.

$$max(-P'lnP-W'lnW) \qquad (۴)$$

تحت (JK) محدودیت گشتاوری،

$$\sum_{i=۱}^{T}y_{ij}x_{ik}=\sum_{i=۱}^{T}x_{ik}p_{ij}+\sum_{i=۱}^{T}\sum_{h=۱}^{H}x_{ik}\nu_h w_{ijh} \qquad (۵)$$

$$k=۱,۲,...,K\ ,\ j=۱,۲,...,J$$

و همچنین با توجه به اینکه p_{ij} و w_{ijh} باید در شرط تابع احتمال بودن صدق کنند این شرایط را داریم،

$$\begin{cases}\sum_{j=۱}^{J}p_{ij}=۱, & i=۱,۲,...,T\\ \sum_{h=۱}^{H}w_{ijh}=۱, & i=۱,۲,...,T\ \ j=۱,۲,...,J\end{cases} \qquad (۶)$$

حال معادله لاگرانژ را به‌صورت زیر تشکیل می‌دهیم،

$$\begin{aligned}L=&-\sum_{i=۱}^{T}\sum_{j=۱}^{J}p_{ij}ln(p_{ij})-\sum_{i=۱}^{T}\sum_{j=۱}^{J}\sum_{h=۱}^{H}w_{ijh}ln(w_{ijh})\\&-\sum_{k=۱}^{K}\sum_{j=۱}^{J}\lambda_{jk}[\sum_{i=۱}^{T}y_{ij}x_{ik}-\sum_{i=۱}^{T}x_{ik}p_{ij}-\sum_{i=۱}^{T}\sum_{h=۱}^{H}x_{ik}\nu_h w_{ijh}]\\&+\sum_{i=۱}^{T}\alpha_i[\sum_{j=۱}^{J}p_{ij}-۱]+\sum_{i=۱}^{T}\sum_{j=۱}^{J}\gamma_{ij}[\sum_{h=۱}^{H}w_{ijh}-۱]=۰ \qquad (۷)\end{aligned}$$

با مشتق‌گیری از معادله (۷) نسبت به p_{ij} رابطه زیر به‌دست‌آمده می‌آید.

$$\frac{\partial l}{\partial p_{ij}}=-lnp_{ij}-۱-\sum_{k=۱}^{K}\lambda_{jk}x_{ik}+\alpha_i=۰ \qquad (۸)$$

$$-lnp_{ij}=۱-\alpha_i+\sum_{k=۱}^{K}\lambda_{jk}x_{ik}$$

در نتیجه،

$$p_{ij}=e^{-(۱-\alpha_i+\sum_{k=۱}^{K}\lambda_{jk}x_{ik})} \qquad (۹)$$

از طرفی با استفاده از محدودیت $\sum_{j=۱}^{J}p_{ij}=۱$ در رابطه (۹) داریم،

$$\sum_{j=۱}^{J}e^{-(۱-\alpha_i+\sum_{k=۱}^{K}\lambda_{jk}x_{ik})}=۱$$

$$e^{\alpha_i-۱}\sum_{j=۱}^{J}e^{-\sum_{k=۱}^{K}\lambda_{jk}x_{ik}}=۱$$

$$e^{\alpha_i-۱}=\frac{۱}{\sum_{j=۱}^{J}e^{-\sum_{k=۱}^{K}\lambda_{jk}x_{ik}}}$$

و برآورد تابع p_{ij} به‌صورت زیر است:

$$\hat{p}_{ij} = \frac{e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik}}}{\sum_{j=1}^{J} e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik}}} \qquad (10)$$

به‌طور مشابه برای به‌دست‌آمده آوردن $\hat{w}_{ijh}$ نیز از معادله لاگرانژ (۷) نسبت به w_{ijh} مشتق می‌گیریم که نتایج،

$$\frac{\partial l}{\partial w_{ijh}} = -ln w_{ijh} - 1 - \sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h + \gamma_{ij} = 0$$

در نتیجه،

$$w_{ijh} = e^{-(1-\gamma_{ij}+\sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h)} \qquad (11)$$

به‌دست‌آمده می‌آیند. با استفاده از محدودیت $\sum_{h=1}^{H} w_{ijh} = 1$ در رابطه (۱۱) داریم،

$$\sum_{h=1}^{H} e^{-(1-\gamma_{ij}+\sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h)} = 1$$

$$e^{-1+\gamma_{ij}} \sum_{h=1}^{H} e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h} = 1$$

$$e^{-1+\gamma_{ij}} = \frac{1}{\sum_{h=1}^{H} e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h}} \qquad (12)$$

و برآورد تابع p_{ij} به‌صورت زیر است:

$$\hat{w}_{ijh} = \frac{e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h}}{\sum_{h=1}^{H} e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h}} \qquad (13)$$

و بنابراین مقدار e_{ij} به‌دست‌آمده می‌آید،

$$e_{ij} = \sum_{h=1}^{H} \hat{w}_{ijh} \nu_h \qquad (14)$$

اگر قرار دهیم،

$$e_j = \nu_j w_j = \begin{pmatrix} \nu_{j1} & 0 & \dots & 0 \\ 0 & \nu_{j2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \nu_{jT} \end{pmatrix} \begin{pmatrix} w_{j1} \\ w_{j2} \\ \vdots \\ w_{jT} \end{pmatrix}$$

که در آن،

$$\sum_{h=1}^{H} \nu_{ijh} w_{ijh} = e_{ij} \;\; i = 1, 2, ..., T \,, \;\; j = 1, 2, ..., J$$

جواب‌های (۱۰) و (۱۳) به فرم نمایی هستند. پس بنابراین $\hat{p}_{ij}$ و $\hat{w}_{ij}$ همیشه مثبت هستند. عوامل نرمال‌سازی ما را مطمئن می‌سازد که $\hat{p}_{ij}$ و $\hat{w}_{ij}$ دارای خواص تابع احتمال هستند و اطلاعاتی در مورد توزیع احتمالات برای پارامترهای مجهول فراهم می‌کند.

پارامترهای مجهول بازنویسی شده $\hat{p}_{ij}$ و $\hat{e}_{ij}$ فقط بر اساس اطلاعات روابط گشتاوری و نه هیچ پذیره اضافه دیگری پایه‌گذاری شده‌اند.

چون λ_jهای لاگرانژ یکتا نیستند یک روش نرمال‌سازی رایج این است که قرار دهیم $\lambda_1 = \beta_1 = 0$ که

$$0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{K \times 1}$$

روش GME نسبت به روش ماکسیمم آنتروپی و درستنمایی ماکسیمم لوژستیک چندجمله‌ای پذیره‌ها و محدودیت‌های کمتری دارد.

در معادله لاگرانژ L (۷) می‌توان قرار داد $\beta_j = -\lambda_j$.

یک انتخاب برای نقاط ν_j می‌تواند $\pm\frac{1}{\sqrt{T}}$ باشد به‌عنوان‌مثال اگر $T = 100$ و $H = 3$ آنگاه

$\nu_{ij} = (-0/1, 0, 0/1)$.

$$\hat{p}_{ij} = \frac{e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik}}}{\sum_{j=1}^{J} e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik}}}$$

اگر قرار دهیم $\beta_{jk} = -\lambda_{jk}$،

$$\hat{p}_{ij} = \frac{e^{\sum_{k=1}^{K} \beta_{jk} x_{ik}}}{\sum_{j=1}^{J} e^{\sum_{k=1}^{K} \beta_{jk} x_{ik}}}$$

و همچنین،

$$\hat{w}_{ijh} = \frac{e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h}}{\sum_{j=1}^{J} e^{-\sum_{k=1}^{K} \lambda_{jk} x_{ik} \nu_h}}$$

برای $\beta_{jk} = -\lambda_{jk}$ به‌دست‌آمده می‌آوریم.

$$\hat{w}_{ijh} = \frac{e^{\sum_{k=1}^{K} \beta_{jk} x_{ik} \nu_h}}{\sum_{j=1}^{J} e^{\sum_{k=1}^{K} \beta_{jk} x_{ik} \nu_h}}$$

و همچنین،

$$\hat{e}_{ij} = \sum_{h=1}^{H} \hat{w}_{ijh} \nu_h$$

می‌توان قرار داد،

$$\hat{y}_{ij} = \hat{p}_{ij} + \hat{e}_{ij}$$

ضرایب رگرسیونی $\beta_j = -\lambda_j$ برای تعیین هر یک از این ضرایب می‌توان از نسبت بخت‌ها کمک گرفت. در رگرسیون لوژستیک معمولی

که

$$ln(\frac{p_{ij}}{p_{i1}}) = -\lambda_j x_i \Rightarrow ln(\frac{p_{ij}}{p_{i1}}) = -\sum_{k=1}^{K} \lambda_{jk} x_{ik} = \sum_{k=1}^{K} \beta_{jk} x_{ik}$$

$$j = 1, 2, ..., J\ ,\ i = 1, 2, ..., T\ ,\ k = 1, 2, ..., K$$

پس JK مجهول داریم و چون $i = 1, 2, ..., T$ پس تعداد T دستگاه J معادله‌ای داریم ([۱۲]).

معیار ارزیابی

در این مطالعه به‌منظور ارزیابی کارایی برآوردگر برای پیش‌بینی از میانگین مربعات خطا استفاده می‌کنیم. که برابر رابطه زیر است:

$$MSE = \frac{1}{T} \sum_{i=1}^{T} (y_i - \hat{y}_i)^2$$

که در آن T تعداد مشاهدات است و $(y_i - \hat{y}_i)^2$ نشان‌دهنده خطای پیش‌بینی است.

۳ مثال کاربردی

در این پژوهش به‌منظور مقایسه روش GME و ML از داده‌های واقعی که مربوط به مشتریان بانک بودند استفاده‌شده است. داده‌های این پژوهش از بانکی در کشور پرتغال گردآوری‌شده است که مربوط به سال ۲۰۱۴ است. جامعه آماری شامل مشتریان بانک است که حجم این داده‌ها حدود ۴۰۰۰ بوده و نمونه‌ای به حجم ۳۹۹ نفر به‌طور تصادفی گرفته‌شده است. بر اساس مبانی نظری مطرح‌شده در مقدمه از متغیرهای توضیحی از قبیل سن (Age)، شغل (Job)، وضعیت تأهل $(Maritalstatus)$، میزان تحصیلات $(Educationlevel)$ و وام (منظور این است که آیا مشتری وام گرفته است یا خیر.) به‌عنوان متغیرهای مؤثر در پیش‌بینی عملکرد مشتریان در راستای افتتاح حساب بلند مدّت استفاده گردید و تقاضای مشتری برای افتتاح حساب بلند مدّت $(termdeposit)$ به‌عنوان متغیر پاسخ گرفته شد. متغیرهایی که کیفی بودند مانند شغل، وضعیت تأهل، میزان تحصیلات و وام به‌صورت متغیرهای ظاهری (۰ و ۱) تعریف شدند.

در اینجا مدل رگرسیونی لوژستیک به‌صورت زیر تعریف می‌شود:

$$Logit(DEP_i) = \beta_0 + \beta_1 AG_i + \beta_2 JO_i + \beta_3 MA_i + \beta_4 ED_i$$

$$+\beta_5 LO_i + \epsilon_i$$

که در آن $DEP_i = 1$ اگر مشتریان افتتاح حساب بلند مدّت داشته باشند و صفر درغیراین صورت. β_i متغیرهای پیش‌گو و ϵ_i جملات خطا هستند. AG_i متغیر ظاهری است، $AG_i = 1$ اگر سن عامل مؤثر برای افتتاح حساب بلند مدّت باشد، درغیراین صورت $AG_i = 0$. شغل JO_i، وضعیت تأهل MA_i، میزان تحصیلات ED_i و وام LO_i وجود دارند.

در این پژوهش از نرم‌افزار گمز برای تحلیل استفاده‌شده است. نتایج برازش مدل رگرسیونی لوژستیک با استفاده از دو روش برآورد GME و ML برای سه حجم نمونه مختلف ۱۸۰، ۲۷۰، ۳۹۹ نفر در جدول (۱) آورده شده است.

طبق جدول (۱) مشاهده می‌شود که در حجم‌های مختلف نمونه، مقدار آماره MSE روش برآورد GME از ML کمتر است و لذا منجر به نتایج دقیق‌تری می‌شود. همچنین مشاهده می‌شود که در حجم‌های نمونه کمتر، روش برآورد GME کارایی بیشتری نسبت به ML دارد به‌طوری‌که تفاوت آماره MSE دو روش برآورد خیلی بیشتر از حجم‌های نمونه بزرگ‌تر است. ظاهراً هر چه حجم نمونه بیشتر شده، مقدار MSE هر دو روش GME و ML کمی افزایش پیداکرده است امّا در حجم نمونه‌های بالاتر دیگر شاهد این افزایش نیستیم و در واقع این مقادیر برای حجم نمونه‌های خیلی بالاتر کاهش می‌یابند.

۴ بحث و نتیجه‌گیری

این تحقیق سعی بر بررسی یک روش جایگزین روش برآورد ML دارد، یعنی روش GME در مدل رگرسیون لوژستیک، که در این تحقیق از داده‌های واقعی موجود در بانکی در کشور پرتغال به‌منظور بررسی احتمال افتتاح حساب بلند مدّت توسط مشتریان این بانک استفاده گردید. بدین منظور از اطلاعاتی از قبیل سن، شغل، وضعیت تأهل، میزان تحصیلات و وام (منظور این است که آیا مشتری وام گرفته است یا خیر.) استفاده گردید. با استفاده از نرم‌افزار گمز، دو روش برآورد GME و ML برای برآورد پارامترهای این مدل به کار گرفته شدند. درنهایت با توجه به مقدار آماره MSE نتیجه گرفته شد که روش برآورد GME دارای نتایج قابل‌اعتمادتری است، پیشنهادشده است که روش GME ممکن است نسبت به روش ML برتری داشته باشد، نتیجه حاصل از این پژوهش با نتیجه گرفته‌شده [۲]، [۹] و [۱۱] کاملاً همسو است.

نام متغیرهای توضیحی	شرح	نوع متغیرهای توضیحی
سن	سن مشتری	عددی
شغل	شغل مشتری	طبقه‌بندی: مدیر کارگر کارآفرین خدمتکار مدیریت بازنشسته آزاد خدماتی دانشجو تکنسین بیکار ناشناخته
وضعیت تأهل	وضعیت تأهل مشتری	طبقه‌بندی: مجرد مطلقه متاهل
تحصیلات	سطح تحصیلات مشتری	طبقه‌بندی: بی‌سواد زیر دیپلم بالاتر از دیپلم ناشناخته
وام	منظور این است که آیا مشتری وام گرفته است یا خیر	طبقه‌بندی: بله خیر

مراجع

[۱] صانعی‌طبس، م. (۱۳۹۴). اصل ماکسیمم آنتروپی تعمیم‌یافته مرتبه α و برآورد پارامترهای مدل رگرسیونی به روش ماکسیمم آنتروپی تعمیم‌یافته. رساله دکتری (آمار ریاضی)، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد.

[2] Ayusuk, A. and Autchariyapanitkul, K. (2017). Factors Influencing Tourism Demand to Revisit Pha Ngan Island Using Generalized Maximum Entropy. *Thai Journal of Mathematics*, 187-196.

جدول ۱: برآورد پارامتر مدل رگرسیون لوژستیک برای افتتاح حساب بلند مدّت

n	۱۸۰		۲۷۰		۳۹۹	
ضرایب	ML	GME	ML	GME	ML	GME
β_0	۳/۴۲۶	۱/۴۲۱	۱/۶۸۴	۰/۴۵۷	۱/۵۰۱	۰/۵۶۰
β_1	−۰/۰۱۹	۰/۰۰۲	−۷/۹۵۴	۰/۰۱۶	−۷/۰۳۶	۰/۰۱۶
β_2	−۰/۴۰۹	−۰/۰۶۸	−۱/۵۰۲	−۰/۰۵۲	−۱/۴۶۲	−۰/۰۲۹
β_3	−۰/۲۶۵	−۰/۳۱۳	۰/۵۹۵	−۰/۲۸۹	−۱/۳۱۰	−۰/۳۱۹
β_4	−۱/۱۵۹	−۰/۰۱۸	−۷/۱۵۱	۰/۰۶۸	۳/۱۷۷	۰/۰۲۱
β_5	−۱/۵۰۰	−۱/۲۲۳	−۱/۳۶۵	−۰/۹۷۰	−۱/۳۴۵	−۱/۱۳۲
MSE	۰/۳۶۱	۰/۱۷۲	۰/۴۲۵	۰/۲۷۳	۰/۴۵۳	۰/۳۹۳

[3] Ciavolino, E. and Calcagni, A. (2016). A Generalized Maximum Entropy (GME) estimation approach to fuzzy regression model. **38**, 51-63.

[4] Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, **41**, 155–160.

[5] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A*, **222**, 309–368.

[6] Golan, A. (2002). Information and Entropy Econometrics Editor's View. *Journal of Econometrics*, **107**, 1-15.

[7] Golan, A. (2008).*Information and Entropy EconometricsA Review and Synthesis.* New York, Now Publishers.

[8] Golan, A., Judge, G. and M.Perloff, J. (1996). A maximum entropy approach to recovering information from multinomial response data. *Journal of the American Statistical Association*, **91**, 841-853.

[9] Golan, A., Moretti, E. and Perloff, J. M. (2001). A SmallSample Estimator for the SampleSelection Model. *CUDARE Working Papers in University of California*, Berkeley.

[10] Judge, G. G. and Golan, A. (1992). Recovering information in the case of ill-posed inverse problems with noise. *Mimo Department Of Agricultural And Natural Resources*, University of California, Berekeley, CA.

[11] Maneejuk, P. (2021). On regularization of generalized maximum entropy for linear models. *Soft Computing*, **25**, 7867–7875.

[12] Mittelhammer, R. C., Judge, G. G. and Miller, D. J. (2000). *Econometric Foundations*. New York, Cambridge University Press.

[13] Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London Series A*, **185**, 71–110.

[14] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series*, **50**, 157–175.

[15] Pukelsheim, F. (1994). The three-sigma Rule. *The American Statistician*, **48**, 88-91.

[16] Song, H. and Witt, S. F. (2000). *Tourism demand modelling and forecasting: Modern econometric approaches*. Routledge.

[17] Sriboonchitta, S., Liu, J. and Sirisrisakulchai. (2015). Willingness-to-pay estimation using generalized maximum-entropy: A case study, *International Journal of Approximate Reasoning*, **60**, 1-7.

[18] Thi Binh An, D., Tsuchida, J. and Yadohisa, H. (2021). K-means generalized maximum entropy estimation for structural equation modeling. *Behaviormetrika*, **48**, 103–115.

Estimation of Logistic Regression Model Parameters Using Generalized Maximum Entropy

Mahsa Markani[1] and Manije Sanei Tabas[2] and Habib Naderi[3] and Hamed Ahmadzadeh[4] and Javad Jamalzadeh[5]

Abstract:

When working with a regression data set, some conditions may not be met and there may be limitations to running the regression model, which makes it difficult for us to use the least squares method. The generalized maximum entropy method with regression infrastructure is able to estimate the parameters of the regression model without considering any conditions on the probability distribution of errors. The capability of this method in small sample sizes has already been investigated and confirmed. When the response variable is a qualitative variable, the logistic regression method is used. In this study, we first introduced the generalized maximum entropy method for a logistic regression model. A random sample of bank customers was collected and in this study to estimate the model parameters from the binary logistic regression model using two generalized maximum entropy methods and maximum likelihood analysis and statistical work was performed and finally compared the two methods. Based on the mean square error statistics for predicting customer demand for long-term account opening, which was obtained from logistic regression using the generalized maximum entropy and maximum likelihood methods, it was found that the generalized maximum entropy estimation method has accurate results.

Keywords: Entropy, Generalized maximum entropy, Logistic regression, Logit, Maximum likelihood.

[1] **Master student, University of Sistan and Baluchestan**

[2] **Assistant Professor, Dept.of statistics, Faculty of Math, University of Sistan and Baluchestan, Daneshgah Ave., Zahedan, Iran.**

[3] **Assistant Professor, Dept.of statistics, Faculty of Math, University of Sistan and Baluchestan, Daneshgah Ave., Zahedan, Iran.**

[4] **Assistant Professor, Dept.of statistics, Faculty of Math, University of Sistan and Baluchestan, Daneshgah Ave., Zahedan, Iran.**

[5] **Assistant Professor, Dept.of statistics, Faculty of Math, University of Sistan and Baluchestan, Daneshgah Ave., Zahedan, Iran.**