

تحلیل بیزی مدل دو متغیره ترتیبی نامتقارن بر پایه متغیر پنهان

رسول قره‌آغاچی اصل^۱، محمد رضا مشکانی^۲، سقراط فقیه‌زاده^۳، انوشیروان کاظم‌نژاد^۴، غلامرضا بابایی^۵، فرید زایری^۶

^۱دانشگاه علوم پزشکی ارومیه، ^۲دانشگاه شهید بهشتی، ^۳دانشگاه تربیت مدرس،

^۴دانشگاه علوم پزشکی شهید بهشتی

تاریخ دریافت: ۱۳۸۶/۹/۲۱ تاریخ آخرین بازنگری: ۱۳۸۶/۱۲/۲۷

چکیده: مدل‌بندی پاسخ‌های ترتیبی همبسته معمولاً پیچیده‌تر از پاسخ‌های پیوسته یا دو حالتی است. روش‌های موجود در برخی حالات، به ویژه وقتی پاسخ دو یا چند متغیره مورد بررسی به صورت نامتقارن باشد، چندان توسعه نیافته‌اند. پیش از این روش‌های مختلفی برای تحلیل پاسخ‌های ترتیبی و همبسته در کتب و مقالات پیشنهاد شده‌اند. در این‌گونه مدل‌بندی‌ها اگر حجم نمونه کم باشد تحلیل کلاسیک کارایی ندارد و بهترین روش فایق آمدن به این مشکل تحلیل مدل با رهیافت بیزی است. در این مقاله روش مدل‌بندی متغیر پنهان با یک توزیع پایه دو متغیره نامتقارن بکار برده و در با رهیافت بیزی تحلیل کرده‌ایم. با استفاده از پیشینهای خاص و الگوریتم MCMC بهره گرفته و پارامترها را برآورد نموده‌ایم. به عنوان کاربرد این مدل را به داده‌های مربوط به زوج متغیرهای رتبه‌ای از چشم‌های راست و چپ ۱۱۶ بیمار رتینوپاتی دیابتی بر حسب تعدادی متغیر مستقل برآزandه‌ایم.

واژه‌های کلیدی: پاسخ‌های ترتیبی نامتقارن، MCMC، متغیر پنهان.

۱ مقدمه

مقیاس‌های ترتیبی زمانی کاربرد پیدا می‌کنند که اندازه‌گیری‌های کمی دقیقی را برای مقدار متغیر در اختیار نداریم و فقط حدود مشخصی برای تفکیک طبقات مشاهدات در نظر گرفته می‌شود. در این تحقیق حالتی را در نظر می‌گیریم که متغیر پاسخ رسته‌ای دو متغیره ترتیبی^۱ است. برای مثال در مطالعات چشم پزشکی، داده‌های مربوط به دو چشم فرد را می‌توان پاسخ‌های دو متغیره همبستگی در نظر گرفت. از ویژگی‌های مهم پاسخ‌های دو متغیره، وجود همبستگی نسبتاً شدید بین زوج پاسخ است. بنابراین در روش‌های تحلیل آماری که برای ارزیابی عوامل مؤثر بر پاسخ‌های زوجی به کار گرفته می‌شوند باید این ساختار همبستگی رعایت شود. مدل‌های خطی تعمیم‌یافته^۲ (GLM) ابزاری مناسب برای تحلیل این‌گونه پاسخ‌ها هستند.

مدل‌های خطی تعمیم‌یافته خانواده بزرگی از مدل‌های آماری هستند که اولین بار توسط نلدر و ودربن (۱۹۷۲) معرفی شدند. به کمک مدل‌بندی آماری می‌توان تأثیر گروهی از عوامل خطر یا کمکی را بر یک یا چند متغیر پاسخ مورد بررسی قرار داد. بسته به نوع متغیر پاسخ تحت بررسی، مدل‌های مختلفی را می‌توان برای تحلیل داده‌ها مورد استفاده قرار داد. به عنوان مثال، وقتی پاسخ از نوع دو حالتی^۳ باشد تابع پیوند را می‌توان توزیع لوژستیک یا نرمال گرفت که در این صورت مدل لوژیت یا پربویت حاصل خواهد شد.

مدل‌بندی پاسخ‌های ترتیبی پیچیده‌تر از پاسخ‌های دو حالتی و پیوسته است. افزون بر این اگر پاسخ‌ها به صورت چندمتغیره همبستگی باشند و ما در مدل‌بندی و تحلیل این همبستگی را رعایت نکنیم، منجر به استنباط‌هایی گمراه کننده خواهد شد. البته در نظر گرفتن این همبستگی مستلزم بکارگیری مدل‌هایی به مراتب پیچیده‌تر از مدل‌هایی است که برای پاسخ‌های یک یا چندمتغیره مستقل به کار می‌روند.

^۱ Bivariate Ordinal

^۲ Generalized Linear Models

^۳ Binary

با مروری بر مطالعات گذشته ملاحظه می‌شود که پژوهشگران برای تحلیل داده‌های پاسخ ترتیبی و بررسی ارتباط آن‌ها با متغیرهای مستقل یا عوامل خطر بیشتر از آزمون‌های آماری استفاده می‌کردند تا مدل‌بندی آماری. در برخی موارد نیز این پاسخ‌های ترتیبی را به پاسخ‌های دو حالتی تبدیل کرده، سپس با کمک مدل‌بندی‌های موجود مورد بررسی قرار می‌دادند. اسنل (۱۹۶۴) و باک و جونز (۱۹۶۸) اولین کسانی بودند که مدل‌های پاسخ رتبه‌ای را مورد توجه و تحلیل قرار دادند. مه‌کولا (۱۹۸۰) یک مدل رگرسیون تجمعی (لوجیت تجمعی) برای پاسخ‌های یک متغیر به صورت

$$\begin{aligned} \text{logit}[P(Y \leq j)] &= \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) \\ &= \log\left(\frac{\pi_1 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_J}\right) \end{aligned}$$

پیشنهاد کرد. با وجود کارایی مدل مه‌کولا برای تحلیل پاسخ‌های ترتیبی یک متغیره، مسئله مدل‌بندی داده‌های چندمتغیره ترتیبی همچنان مسئله‌ای پیچیده باقی ماند. سال‌ها بعد آمارشناسان برای حل این مشکل سه راهبرد پیشنهاد کرده‌اند که عبارتند از:

مدل‌های حاشیه‌ای^۴: روسنر (۱۹۸۴) دو مدل مناسب برای تحلیل پاسخ‌های دو متغیره با توزیع دوجمله‌ای و نرمال ارائه کرد. این دو مدل رگرسیون چندگانه، ابزارهایی مناسب برای تحلیل پاسخ‌های چندمتغیره در اختیار تحلیل‌گران قرار می‌دادند که به کمک آن‌ها همبستگی بین پاسخ‌ها در مدل گنجانده می‌شد. روشن معادلات برآورده لیانگ و زیگر (۱۹۸۶) را می‌توان مهم ترین تحول در زمینه تحلیل داده‌های چندمتغیره همبسته نامید. این روش، بعدها به نام روش معادلات برآورده تعمیم یافته^۵ مشهور شد. این روش بطور مستقیم برای تحلیل پاسخ‌های ترتیبی همبسته قابل استفاده نبود، اما لیپشتز و همکاران (۱۹۹۴) این روش را برای مدل‌بندی داده‌های رسته‌ای همبسته تعمیم دادند.

^۴ Marginal models

^۵ Generalized Estimating Equations

مدل‌های اثرات تصادفی^۶: این مدل بر اساس جملات تصادفی مربوط به خوشه بنا شده و در آن پارامترها تفسیری شرطی خواهند داشت. در یک مدل حاشیه‌ای برای تمامی افراد مورد بررسی ارتباط یکسانی بین متغیر پاسخ و ماتریس متغیرهای کمکی فرض می‌شود. در مقابل در مدل اثرات تصادفی این امکان هست که نوع ارتباط بین متغیر پاسخ و ماتریس متغیرهای کمکی برای افراد مختلف متفاوت باشد. جملات اثرات تصادفی غیرقابل مشاهده‌اند، اما از توزیعی مشخص پیروی می‌کنند.

مدل متغیر پنهان^۷: اولین کسی بود که یک مدل را با فرض یک متغیر پنهان برای داده‌های ترتیبی به کار برد. پس از وی، مه‌کولا (۱۹۸۰) مدل رگرسیونی را برای تحلیل پاسخ‌های یک متغیره رسته‌ای با فرض یک متغیر پنهان پیوسته که به طور مستقیم قابل مشاهده نیست ارائه کرد. کیم (۱۹۹۵) مدل دومتغیره پنهان را بر پایه توزیع دومتغیره نرمال برای پاسخ‌های دومتغیره ترتیبی به کار برد. آلبرت و چیب (۱۹۹۳) از تحلیل بیزی برای پاسخ‌های یک متغیره دو حالت و چند حالتی با مدل متغیر پنهان استفاده کردند، چیب و گرینبرگ (۱۹۹۸) مدل چندمتغیره رسته‌ای را با روش مدل متغیر پنهان و با فرض توزیع پایه نرمال به روش بیزی تحلیل کرده‌اند. ویلیامسون و همکاران (۱۹۹۵) برای یک مدل رتبه‌ای یک متغیره با فرض توزیع پنهان نرمال (مدل پروبیت) از تحلیل بیزی استفاده نموده و پارامترهای مدل را برآورد نمودند. اُبرین و دانسون (۲۰۰۴) برای مدل رتبه‌ای چندمتغیره و با روش متغیر پنهان مبتنی بر توزیع لوزیستیک از روش بیزی استفاده کرده و پارامترهای مدل را برآورد کردند. بیسوس و داس (۲۰۰۲) مدل کیم را به روش بیزی تحلیل کرده و پارامترهای مدل را برآورد کردند. زایری و کاظم‌نژاد (۲۰۰۶) مدلی برای پاسخ‌های دومتغیره رسته‌ای نامتقارن با استفاده از توزیع تجمعی گامبل تعمیم یافته ارائه نمودند و پارامترهای مدل را با رهیافت کلاسیک برآورد کرده و برآزش بهتر این مدل را نسبت به مدل نامتقارن نشان داده‌اند. اما این تحلیل به دلیل دومتغیره بودن و تعدد سطوح رسته‌های پاسخ برای حجم نمونه کم (بسته به تعداد سطوح پاسخ حتی بیشتر از ۵۰ نمونه) کارا نیست و با پیش فرض‌های تحلیل

^۶ Random Effect Models

^۷ Latent Variable Models

کلاسیک مطابقت ندارد، چون تحلیل‌های کلاسیک مبتنی بر نمونه بزرگ و توزیع‌های مجانبی هستند (اگرستی، ۲۰۰۲). ما در این مقاله برای رفع این نقصان، مدل رگرسیونی پاسخ‌های دو متغیره ترتیبی همبسته را با روش مدل متغیر پنهان مبتنی بر توزیع تعمیم‌یافته دو متغیره نامتقارن گامبل (سترثویت و هاچینسون، ۱۹۷۸)، با رهیافت بیزی تحلیل کرده و پارامترهای مدل را با الگوریتم‌های مونت کارلوی زنجیر مارکوفی^۸ (MCMC) برآورد می‌کنیم. در بخش ۲ مدل دو متغیره نامتقارن معرفی می‌شود و توزیع‌های شرطی کامل پارامترها به دست آورده می‌شوند. در بخش ۳ داده‌های رتینوپاتی بیماران دیابتی به عنوان کاربرد مدل به روش بیزی تحلیل شده است. در نهایت در بخش ۴ بحث و نتیجه‌گیری به عمل آمده است.

۲ مدل نامتقارن

در مدل‌هایی که در بالا معرفی شدند، اگر فرض بر این باشد که پاسخ‌ها نامتقارن هستند، یعنی گرایش احتمال به سمت صفر و یک یکسان باشد، تابع پیوند یا توزیع متغیر پنهان را نرمال یا لوژستیک می‌گیرند. مدل‌های حاصل به مدل پربویت تجمعی یا لوجیت تجمعی معروفند. اما در اکثر موقعیت‌فرض (تقارن) صادق نیست و لازم است که تابع پیوند یا توزیع متغیر پنهان نامتقارن انتخاب شود. در حالت یک متغیره اگر توزیع یک متغیره گامبل را که یک توزیع نامتقارن است به عنوان تابع پیوند مدل در نظر بگیریم، مدل حاصل به مدل لگ-لگ مکمل معروف است. گامبل (۱۹۶۱) تابع توزیع دو متغیره نامتقارنی ارائه کرد، که می‌توان از آن برای پاسخ دو متغیره نامتقارن استفاده کرد. تابع توزیع دو متغیره گامبل به شکل

$$\Psi(x, y) = (1 + e^{-x} + e^{-y})^{-1} \quad -\infty < x, y < +\infty$$

است، که اشکال بزرگ آن فقدان پارامتری برای بیان همبستگی بین دو متغیر تصادفی X و Y است. سترثویت و هاچینسون (۱۹۷۸) توزیع فوق را به صورت

$$\Psi(x, y) = (1 + e^{-x} + e^{-y})^{-v}, v > 0, -\infty < x, y < +\infty$$

^۸ Markov Chain Monte Carlo

۱۴۴ تحلیل بیزی مدل دو متغیره ترتیبی نامتقارن بر پایه متغیر پنهان

باتابع چگالی

$$f(x, y) = \frac{v(v+1)e^{-x}e^{-y}}{(1+e^{-x}+e^{-y})^{v+2}}, \quad v > 0, \quad -\infty < x, y < +\infty \quad (1)$$

اصلاح کردند، که در آن v پارامتری است که همبستگی بین X و Y را توصیف می‌کند (برای جزئیات بیشتر درباره این توزیع و خواص مهم آن نظریه: توزیع‌های حاشیه‌ای، امید ریاضی، واریانس و کواریانس X و Y گامبل، ۱۹۶۱ و سترثویت و هاچینسون، ۱۹۷۸ را مشاهده کنید). در این توزیع دو متغیره ضریب همبستگی بین X و Y را می‌توان از رابطه

$$\rho = \frac{\zeta(2, v)}{\zeta(2, v) + \pi^2/6} \quad (2)$$

محاسبه نمود، که در آن $\zeta(s, a) = \sum_{m=0}^{\infty} (m+a)^{-s}$ تابع زتا ریمان^۹ است. با محاسباتی سرراست می‌توان نشان داد که در این توزیع $1 \leq \rho \leq 0$ است.

۱.۲ مدل دو متغیره ترتیبی نامتقارن

فرض کنید (Y_{1i}, Y_{2i}) نشان‌دهنده پاسخ دو متغیره رتبه‌ای برای یک فرد باشد، مانند وضع چشم راست و چشم چپ که می‌توانند رتبه‌های $k = 1, \dots, n$ را انتخاب کنند. همچنین فرض کنید (Y_{1i}^*, Y_{2i}^*) متغیرهای پنهان متناظر با (Y_{1i}, Y_{2i}) به صورت

$$\theta_{jk-1} < y_{ji}^* \leq \theta_{jk} \quad \text{if } y_{ji} = k, \quad i = 1, \dots, n$$

باشند، که در آن θ ‌ها را مجموعه نقاط آستانه‌ای (برش) می‌نامیم و به صورت

$$\theta_j = (\theta_{j0}, \theta_{j1}, \dots, \theta_{jk}), \quad \theta_{j0} = -\infty, \quad \theta_{jk} = +\infty$$

تعریف می‌شوند. پس زوج متغیر (y_{1i}, y_{2i}) برای شخص i می‌تواند مقدار (h, ℓ) را بگیرد و فضای نمونه $S = \{(h, \ell) : h, \ell = 0, 1, \dots, k\}$ است. با این فرضیات

^۹ Riemann Zeta function

قره آغاجی، مشکانی، فقیهزاده، کاظم نژاد، بابایی، زایری ۱۴۵

مدل رگرسیونی دو متغیره برای (Y_{1i}^*, Y_{2i}^*) به صورت $Y_i^* = X_i' \beta + \epsilon_i$ است، که در آن X_i' بردار متغیرهای کمکی فرد i است و

$$\epsilon_i = (\epsilon_{1i}, \epsilon_{2i})' \sim SH(x_i' \beta, 1) \quad i = 1, \dots, n$$

از نوع توزیع دو متغیره سترشویت و هاچینسون (SH) مندرج در (۱) است.

بردار پاسخ عبارت است از $Y = (Y_1', \dots, Y_i', \dots, Y_n')$ که در آن $Y_i = (Y_{1i}', Y_{2i}')$ و به طور مشابه بردار متغیر پنهان نیز $(Y_{1i}^{*'}, \dots, Y_{ni}^{*'})$ است، که در آن $Y^* = (Y_{1i}^{*'}, \dots, Y_{ni}^{*'})$ است. می باشد. بردار $\beta = (\beta_1', \beta_2')$ ضرایب مدل رگرسیونی است. درستنایی کامل Y و Y^* نیز به صورت

$$f(y, y^* | \beta, \theta, v) = \prod_{i=1}^n \left[\sum_{h, \ell \in S} I_{h\ell}^i I(\theta_{1h-1} < y_{1i}^* \leq \theta_{1h}, \theta_{2\ell-1} < y_{2i}^* \leq \theta_{2\ell}) \right] \times \frac{v(v+1)e^{-(y_{1i}^* - x_{1i}' \beta_1)} e^{-(y_{2i}^* - x_{2i}' \beta_2)}}{(1 + e^{-(y_{1i}^* - x_{1i}' \beta_1)} + e^{-(y_{2i}^* - x_{2i}' \beta_2)})^{v+2}}$$

خواهد بود، که در آن $(h, \ell) \in S$ و

$$I_{h\ell}^i = \begin{cases} 1 & y_{1i} = h, y_{2i} = \ell \\ 0 & \text{جاهای دیگر} \end{cases}$$

می باشند. در صورتی که β های جداگانه برای اندام اول و دوم (چشم چپ و راست) فرض شود $\beta_1 \neq \beta_2$ ماتریس مشاهدات متغیرهای کمکی به صورت

$$X_i = \begin{pmatrix} x_{1i}' & \circ' \\ \circ' & x_{2i}' \end{pmatrix}$$

خواهد بود. اما اگر $\beta_1 = \beta_2 = \beta$ فرض شود، یعنی ضرایب متغیرهای کمکی مدل یکسان فرض شوند، ماتریس طرح مدل به صورت

$$X_i = \begin{pmatrix} x_i' & x_{1i}' \\ x_i' & x_{2i}' \end{pmatrix}$$

در خواهد آمد، که در آن x_i' ها متغیرهای مربوط به فرد i هستند که برای هر دو چشم یکسان اند و x_{ji}' ها متغیرهای مربوط به هر اندام اند که برای چشم راست و

چپ متفاوت هستند. توزیع توأم پیشینی پارامترها به صورت زیر است:

$$\begin{aligned}\pi(\underline{y}^*, \underline{\beta}, \underline{\theta}, v | y, D) &= \pi(\underline{y}; \underline{y}^* | \underline{\beta}, \underline{\theta}, v) \times \pi(\underline{\beta}, \underline{\theta}, v) \\ &\propto I(\underline{y}; \underline{y}^*, \underline{\theta}) \times \pi(\underline{y}^* | \underline{\beta}, v) \times \pi(\underline{\beta}) \pi(\underline{\theta}) \pi(v) \\ -\infty < y^*, \beta, \theta < +\infty, v > .\end{aligned}\quad (3)$$

۲.۲ توزیع‌های پیشینی پارامترها

با فرض نامتقارن بودن پاسخ‌ها، توزیع متغیر نهفته Y_i^* را توزیع دومتغیره سترثویت و هاچینسون (۱) انتخاب می‌کنیم، که تابع پیوند مدل نیز می‌باشد. توزیع پیشینی β را نرمال چندمتغیره $N_p(\mu, \Sigma_0)$ در نظر می‌گیریم. برای توزیع پیشینی θ یک توزیع ناآگاهی بخشن $1 = (\theta)^\pi$ را در نظر می‌گیریم. برای v که پارامتر همبستگی است، رفتار این پارامتر و رابطه آن با ρ یعنی رابطه (۲) را بررسی می‌کنیم. می‌توان نشان داد که اگر $v \rightarrow \infty$ آنگاه $\rightarrow \infty$ و در نتیجه $\rho \rightarrow 0$ ، همچنان اگر $v \rightarrow -\infty$ آنگاه $\rho \rightarrow 0$ و متعاقب آن $v \rightarrow \infty$. با این تفاسیر و با توجه به این که پاسخ‌های دومتغیره ما دارای همبستگی نسبتاً شدیدی هستند، باید v دارای توزیعی چوله به راست با مقدار مدبین صفر و یک باشد. پس توزیع $\Gamma(a, b)$ توزیع پیشینی مناسب برای این پارامتر می‌تواند باشد.

۳.۲ توزیع‌های شرطی کامل

در مدل رگرسیون پنهان دومتغیره رتبه‌ای نامتقارن چهار گره تصادفی Y^* و θ و β و v وجود دارند. بنابراین در هر مرحله از شبیه‌سازی MCMC لازم است چهار توزیع شرطی کامل متناظر با هر کدام از گره‌های فوق روزآمد شده و نمونه‌گیری شوند. توزیع شرطی کامل پارامتر Y^* : با توجه به توزیع پیشینی توأم پارامترها (۳)، توزیع شرطی کامل Y^* به شکل

$$\pi\left(y_i^* | \beta, \theta, v, y_i = \binom{h}{\ell}, D\right)$$

قره آغاچی، مشکانی، فقیهزاده، کاظم نژاد، بابایی، زایری ۱۴۷

$$\begin{aligned} & \propto \sum_{(h,\ell) \in S} I_{h\ell}^i I \left(\begin{array}{l} \theta_{1h-1} < y_{1i}^* \leq \theta_{1h} \\ \theta_{2\ell-1} < y_{2i}^* \leq \theta_{2\ell} \end{array} \right) \frac{v(v+1)e^{-(y_{1i}^*-x'_{1i}\beta_1)}e^{-(y_{2i}^*-x'_{2i}\beta_2)}}{[1+e^{-(y_{1i}^*-x'_{1i}\beta_1)}+e^{-(y_{2i}^*-x'_{2i}\beta_2)}]^{v+2}} \\ & \propto \sum_{(h,\ell) \in S} I \left(\begin{array}{l} \theta_{1h-1} < y_{1i}^* \leq \theta_{1h} \\ \theta_{2\ell-1} < y_{2i}^* \leq \theta_{2\ell} \end{array} \right) \frac{e^{-(y_{1i}^*+y_{2i}^*)}}{[1+e^{-(y_{1i}^*-x'_{1i}\beta_1)}+e^{-(y_{2i}^*-x'_{2i}\beta_2)}]^{v+2}} \end{aligned}$$

در می آید، که در آن $I_{h\ell}^i = I^i(y_{1i} = h, y_{2i} = \ell)$ می باشد. این توزیع شرطی کامل، یک توزیع دو متغیره سترشویت و هاچینسون بریده شده در دو ناحیه بالا است.

توزیع شرطی کامل θ : توزیع شرطی کامل θ نیز به صورت

$$\begin{aligned} & \pi \left[\begin{pmatrix} \theta_{1h} \\ \theta_{2\ell} \end{pmatrix} | \beta, v, \theta_{\binom{n}{2}}, y^*, y, D \right] \\ & \propto \prod_{i=1}^n \left[I_{h,\ell}^i I \left(\begin{array}{l} \theta_{1h-1} < y_{1i}^* \leq \theta_{1h} \\ \theta_{2\ell-1} < y_{2i}^* \leq \theta_{2\ell} \end{array} \right) + I_{h+1,\ell}^i I \left(\begin{array}{l} \theta_{1h} < y_{1i}^* \leq \theta_{1h+1} \\ \theta_{2\ell-1} < y_{2i}^* \leq \theta_{2\ell} \end{array} \right) \right. \\ & \quad \left. + I_{h,\ell+1}^i I \left(\begin{array}{l} \theta_{1h-1} < y_{1i}^* \leq \theta_{1h} \\ \theta_{2\ell} < y_{2i}^* \leq \theta_{2\ell+1} \end{array} \right) + I_{h+1,\ell+1}^i I \left(\begin{array}{l} \theta_{1h} < y_{1i}^* \leq \theta_{1h+1} \\ \theta_{2\ell} < y_{2i}^* \leq \theta_{2\ell+1} \end{array} \right) \right] \end{aligned}$$

است، که یک توزیع یکنواخت $\theta_{jc} \sim \text{unif}(t_{\theta c}, r_{\theta c})$ است، که در آن

$$t_{\theta c} = \max \left[\max_{i=1, \dots, n} (y_{ji}^* | y_{ji} = c), \theta_{jc-1} \right]$$

$$r_{\theta c} = \min \left[\min_{i=1, \dots, n} (y_{ji}^* | y_{ji} = c+1), \theta_{jc+1} \right]$$

می باشند، بطور یک

$$y_{ji} = c \Leftrightarrow \theta_{jc-1} < y_{ji}^* < \theta_{jc}.$$

توزیع شرطی کامل پارامتر v : با در نظر گرفتن پیشین $\Gamma(\alpha, \lambda) \sim v$ و با توجه به توزیع پسینی توأم پارامترها (3) ، توزیع شرطی کامل v بصورت

$$\begin{aligned} & \pi(v | \beta, \theta, y^*, y, D) \\ & = \prod_{i=1}^n \frac{v(v+1)e^{-(y_{1i}^*-x'_{1i}\beta_1)}e^{-(y_{2i}^*-x'_{2i}\beta_2)}}{[1+e^{-(y_{1i}^*-x'_{1i}\beta_1)}+e^{-(y_{2i}^*-x'_{2i}\beta_2)}]^{v+2}} \times \frac{\lambda^\alpha v^{\alpha-1} e^{-\lambda v}}{\Gamma(\alpha)} \\ & \propto v^{n+\alpha-1} (v+1)^n e^{\left[-v(\lambda + \sum_{i=1}^n \ln(1+e^{-(y_{1i}^*-x'_{1i}\beta_1)}+e^{-(y_{2i}^*-x'_{2i}\beta_2)}) \right]} \\ & \propto v^{n+1} \Gamma \left(n + \alpha, [\lambda + \sum_{i=1}^n \ln(1+e^{-(y_{1i}^*-x'_{1i}\beta_1)}+e^{-(y_{2i}^*-x'_{2i}\beta_2)})] \right) \end{aligned}$$

محاسبه می‌شود، که دارای شکل شناخته شده‌ای نیست ولی مضربی از یک توزیع گاما است. بنابراین برای نمونه‌گیری از چگالی شرطی v از الگوریتم متروپولیس هستینگز با توزیع پیشنهادی

$$\Gamma(n + \alpha, \lambda + \sum_{i=1}^n \ln[1 + e^{-(y_{1i}^* - x'_{1i}\beta_1)} + e^{-(y_{2i}^* - x'_{2i}\beta_2)}])$$

استفاده نموده و در هر مرحله نمونه تولید کرده و توزیع شرطی کامل v را روزآمد می‌کنیم.

توزیع شرطی کامل پارامتر β : با در نظر گرفتن توزیع پیشینی نرمال برای β ، توزیع شرطی کامل β به صورت

$$\begin{aligned} & \pi(\beta|y^*, \theta, v, y, D) \\ & \propto e^{-\frac{1}{2}(\beta-\mu)' \Sigma_{\circ}^{-1} (\beta-\mu)} \\ & \quad \times \prod_{i=1}^n \frac{v(v+1)e^{-(y_{1i}^* - x'_{1i}\beta_1)}e^{-(y_{2i}^* - x'_{2i}\beta_2)}}{[1 + e^{-(y_{1i}^* - x'_{1i}\beta_1)} + e^{-(y_{2i}^* - x'_{2i}\beta_2)}]^{v+2}} (\frac{1}{2\pi|\Sigma_{\circ}|})^{\frac{n}{2}} \\ & \propto e^{-\frac{1}{2}(\beta-\mu)' \Sigma_{\circ}^{-1} (\beta-\mu)} \\ & \quad \times e^{\sum_{i=1}^n [x'_{1i}\beta_1 + x'_{2i}\beta_2 - (v+2) \ln(1 + e^{-(y_{1i}^* - x'_{1i}\beta_1)} + e^{-(y_{2i}^* - x'_{2i}\beta_2)})]} \end{aligned}$$

به دست می‌آید، که شکل شناخته شده‌ای ندارد و برای نمونه‌گیری از چگالی شرطی β می‌توانیم از الگوریتم متروپولیس هستینگز با توزیع پیشنهادی نرمال p متغیره با میانگینی برابر با مد توزیع شرطی کامل و ماتریس واریانس کوواریانس برابر با وارون ماتریس اطلاع فیشر (هسی) یعنی $[I(\hat{\theta})]^{-1}$ استفاده کنیم، که در آن $I(\theta) = -\frac{\partial^2}{\partial \theta^2} \ln P(\theta|any)$. مد توزیع شرطی کامل را می‌توان از برابر صفر قرار دادن مشتق اول لگاریتم توزیع شرطی کامل و حل عددی این معادله با روش‌هایی چون گاوس-نیوتون به دست آورد. لگاریتم توزیع شرطی کامل به صورت

$$\begin{aligned} \ln \pi(\beta|y^*, \theta, v, y, D) &= Const. - \frac{1}{2}(\beta-\mu)' \Sigma_{\circ}^{-1} (\beta-\mu) \\ &\quad + \sum_{i=1}^n [X'_i \beta - (v+2) \ln(1 + e^{-(Y_i^* - X'_i \beta)})] \end{aligned}$$

و مشتق مرتبه اول آن به صورت

$$\frac{\partial}{\partial \beta} \ln \pi(\beta|y^*, \theta, v, y, D) = -(\beta - \mu)' \Sigma_{\circ}^{-1} + \sum_{i=1}^n [X'_i - (v + 2)] \frac{X'_i e^{-(Y_i^* - X'_i \beta)}}{1 + e^{-(Y_i^* - X'_i \beta)}}$$

است. اینک بردار β ، مدل چگالی شرطی، را از حل معادله بالا با روش عددی گاووس-نیوتون به دست می‌آوریم. مشتق دوم نیز به صورت

$$\frac{\partial^2}{\partial \beta \partial \beta'} \ln \pi(\beta|y^*, \theta, v, y, D) = -\Sigma_{\circ}^{-1} - (v + 2) \sum_{i=1}^n X_i X'_i \frac{e^{-(Y_i^* - X'_i \beta)}}{(1 + e^{-(Y_i^* - X'_i \beta)})^2}$$

است. پس می‌توان با توزیع پیشنهادی نرمال p متغیره

$$N\left(\hat{\beta}, (\Sigma_{\circ}^{-1} + (v + 2) \sum_{i=1}^n X_i X'_i \frac{e^{-(Y_i^* - X'_i \beta)}}{[1 + e^{-(Y_i^* - X'_i \beta)}]^2})^{-1}\right)$$

از الگوریتم متروپولیس هستینگز از چگالی شرطی β نمونه‌گیری کرد.

۳ مثال کاربردی

بیماری رتینوپاتی دیابتی یکی از شایع‌ترین عوارض دیابت (مرض قند) است. این بیماری در دیابتی‌های نوع ۱ بیشتر بروز می‌کند و در صورت تشخیص دیر هنگام باعث مشکلات بینایی و حتی نابینایی می‌شود. در این مثال ۱۱۶ بیمار دیابتی نوع ۱ را که در سن کمتر از ۳۰ سال به دیابت مبتلا شده بودند^{۱۰} از بیمارستان چشم‌پزشکی فارابی به تصادف انتخاب کردیم. متغیر پاسخ شدت رتینوپاتی به وسیله معاینه هر دو چشم با مردمک کاملاً باز توسط وسایل و روش‌های استاندارد به چهار رسته زیر تشخیص داده شدند: ۱- سالم^{۱۱} ۲- خفیف^{۱۲} ۳- متوسط تا شدید^{۱۳} ۴- گسترنده سریع (پرولیفراتیو^{۱۴}). متغیر وابسته (شدت رتینوپاتی جفت چشم بیمار) یک پاسخ دومتغیره همبسته ترتیبی است. متغیرهای کمکی (مستقل)

^{۱۰} Younger Onset

^{۱۱} None

^{۱۲} Mild

^{۱۳} Moderate and Severe

^{۱۴} Proliferative

طول مدت بیماری دیابت، سن در زمان ابتلا، فشار خون سیستولیک، فشار خون دیاستولیک، جنس، شاخص توده بدن، تعداد ضربان نبض، دوز انسولین روزانه، محل سکونت، پروتینوری، عیوب انکساری، فشار داخل چشم و ادم ماکولای نیز به وسیله معاینه یا پرسش از بیمار و همراه بیمار یا مطالعه پرونده بیمار اندازه‌گیری و ثبت گردید. ده متغیر اول متغیرهای شخصی بوده و برای هر دو چشم یکسان هستند و سه متغیر آخر متغیرهای اندامی اند که برای چشم چپ و راست ممکن است مقادیر متفاوت داشته باشند. توزیع توأم شدت رتینوپاتی بر حسب نوع چشم در این بیماران در جدول ۱ آمده است. این جدول نشان می‌دهد که حتی بدون در نظر گرفتن متغیرهای کمکی (مستقل) به دلیل اندازه کوچک نمونه پیش‌فرضهای استنباط کلاسیک برقرار نیست، بنابراین از استنباط بیزی برای تحلیل آن استفاده می‌کنیم. در تحلیل بیزی، ابتدا در مرحله اول به تعداد مشاهدات ۱۱۶ جفت عدد و با توجه به رتبه‌های بدست آمده از بیماران، از توزیع شرطی کامل γ نمونه‌گیری کردیم و در تمامی مراحل بعد نیز باید این تعداد نمونه تولید و روزآمد کنیم. برای نقاط برش (آستانه‌ای) با در نظر گرفتن مقدارهای اولیه $4, 5$ و 6 از توزیع یکنواخت بدست آمده و با الگوریتم نمونه‌گیری گیز داده شبیه‌سازی کردیم. برای نمونه‌گیری از توزیع‌های شرطی β و γ از الگوریتم متropolیس-هستینگز بهره جستیم. مقدار اولیه پارامتر γ را $5/0$ در نظر گرفتیم و با توجه به این که توزیع پیشینی گاما با توجه به رفتار آن تقریباً دارای مد $5/0$ و واریانس $5/0$ است، با حل دو معادله مربوط به فرمول مد و واریانس توزیع گاما، توزیع پیشینی $(\Gamma(2, 2))$ را بدست آوردیم. با اطلاعاتی که در اختیار بود نیازی به تحلیل بیزی سلسله مراتبی پیدا نشد. مقدار β (میانگین توزیع پیشنهادی) را نیز از روش تکراری گاووس-نیوتون به دست آوردیم. توزیع پیشینی پارامتر β را از توزیع نرمال با میانگین صفر و واریانس 1000 که پیشینی مبهم است، استفاده کردیم. برای انجام تحلیل بیزی به روش مونت کارلوی زنجیر مارکوفی از تابع‌های چگالی شرطی پارامترهای طور متوالی نمونه‌گیری کردیم.

۱.۳ اجرای شبیه‌سازی

اجرای شبیه‌سازی توسط نرم‌افزار $R^2.5$ برنامه نویسی شد، به دلیل دیر همگرا شدن پارامتر θ نقاط آستانه‌ای، ۲۰۰۰۰ نمونه از هر توزیع تولید کردیم که ۱۰۰۰۰ نمونه را به عنوان مقادیر دوره تطبیق از نتایج کنار گذاشتیم. جهت محاسبه شاخص‌های مرکزی، پراکندگی و صدک‌های مورد نیاز و انجام تشخیص همگرایی بوسیله رسم نمودار و انجام آزمون همگرایی نمونه‌های به دست آمده توسط BOA (یکی از بسته های نرم‌افزار R) فراخوانی شدند، برای اطمینان نسبت به همگرایی زنجیر از آزمون گه‌ویکه 15 استفاده نمودیم. در این مقاله با فرض $\beta_1 = \beta_2 = v$ نتایج را در جدول ۲ آورده‌ایم. برآورد $v = 0.69$ را در رابطه زتای ریمان (۲) قرار می‌دهیم و همبستگی 0.64 بین دو متغیر پاسخ (شدت رتینوپاتی چشم راست و چپ) بدست می‌آید که همبستگی بالایی است.

۴ بحث و نتیجه‌گیری

مدل‌های متغیر پنهان دو متغیره یکی از بهترین انتخاب‌ها برای تحلیل پاسخ‌های دو متغیره ترتیبی همبسته هستند، چرا که به کمک آنها می‌توان برآوردهای صریح از همبستگی بین مشاهدات پاسخ ترتیبی بدست آمده در حضور انواع متغیرهای کمکی بدست آورده‌اما در دو مدل حاشیه‌ای و متغیر تصادفی چنین نیست. در مورد پاسخ‌های زوجی استفاده از این مدل‌ها بر مدل‌های یک متغیره ارجحیت دارد. در مدل‌های دو متغیره پنهان، پارامترها را می‌توان علاوه بر تفسیرهای معمول (مانند معنی‌داری) بر حسب مقیاس متغیر نهفته نیز تفسیر کرد. این نوع تفسیر، استنباط‌های جالبی در مورد متغیرهای کمکی مختلف و تلفیق آنها بدست می‌دهد. در مدل متغیر پنهان تفسیر ضرایب رگرسیونی برآورد شده را می‌توان بر حسب مقیاس متغیر پنهان در حضور متغیرهای کمکی (اثرگذار) انجام داد.

به عنوان مثال در این مدل و با در نظر گرفتن نتایج جدول ۲، با توجه به ضریب (تأثیر) متغیر طول بیماری $\hat{\beta}_1 = 0.1126$ و وجود آدم ماکولای

جدول ۱: توزیع توانم شدت رتینوپاتی بر حسب نوع چشم.

کل	پرولیفراطیو	متوجه	خفیف	سالم	چشم چپ
					چشم راست
۴۹(%۴۲, ۲۴)	۰	۰	۱۲	۳۷	سالم
۳۶(%۲۱, ۰۳)	۰	۴	۱۸	۱۴	خفیف
۲۰(%۱۷, ۲۴)	۱	۱۲	۵	۲	متوجه
۱۱(%۹, ۴۸)	۹	۲	۰	۰	پرولیفراطیو
۱۱۶	۱۰(%۸, ۶)	۱۸(%۱۵, ۵۱)	۲۵(%۳۰, ۲)	۵۲(%۴۵, ۶)	کل

جدول ۲: نتایج شبیه سازی β (ضرایب متغیرهای کمکی)، v و θ .

% ۹۷, ۵	صدک % ۵۰	صدک % ۲, ۵	صدک % ۰, ۵	انحراف معیار	میانگین	پارامتر
۰, ۱۲۷	۰, ۱۱۲۷	۰, ۰۹۷۸	۰, ۰۰۷۵	۰, ۱۱۲۶	طول مدت بیماری	
-۰, ۰۰۳	-۰, ۰۲۱	-۰, ۰۳۹	۰, ۰۰۹	-۰, ۰۰۲۱	سن در زمان ابتلاء	
۰, ۰۴۴	۰, ۰۲۷	۰, ۰۱۱۶	۰, ۰۰۸	۰, ۰۰۲۸	فشار خون دیاستول	
-۰, ۰۰۸	-۰, ۰۱۶۸	-۰, ۰۲۶	۰, ۰۰۴۶	-۰, ۰۰۱۷	فشار خون سیستول	
۰, ۰۹۷	۰, ۰۶۷	۰, ۰۳۶	۰, ۰۱۶	۰, ۰۰۶۷	شاخص توده بدن	
۰, ۰۱	-۰, ۰۰۳	-۰, ۰۱۷	۰, ۰۰۷	-۰, ۰۰۰۳	تعداد بیض	
-۰, ۰۹۸	-۰, ۰۲۹	-۰, ۰۴۸۸	۰, ۱	-۰, ۰۲۹	جنس	
۰, ۹۸	۰, ۷۶	۰, ۵۳	۰, ۱۱۵	۰, ۷۶	پروتئینوری	
۰, ۰۴	۰, ۰۲۸	۰, ۰۱۴	۰, ۰۰۷	۰, ۰۰۲۸	دوز انسولین روزانه	
۰, ۱۹۸	-۰, ۰۳۶	-۰, ۰۲۷	۰, ۱۲	-۰, ۰۰۳۵	محل سکونت	
۱, ۹	۱, ۶۷	۱, ۴۵	۰, ۱۱۵	۱, ۶۸	ام ماکولای	
۰, ۰۲۹	-۰, ۰۰۵	-۰, ۰۱۳	۰, ۰۴	-۰, ۰۰۵	عیوب انکساری	
۰, ۰۳۹	۰, ۰۱۲۹	-۰, ۰۱۲۹	۰, ۰۱۲	۰, ۰۱۳	فشار داخل چشم	
۵, ۳	۴, ۷۶	۴, ۲۳۶	۰, ۲۷۹	۴, ۷۶	۰, ۱۱	
۷, ۱۲	۶, ۰۵	۶, ۰۰۶	۰, ۲۸۸	۶, ۰۶	۰, ۱۲	
۱۰, ۱۸	۹, ۵۴	۸, ۹	۰, ۳۲۵	۹, ۵۴	۰, ۱۲	
۵, ۵۲	۴, ۹۳	۴, ۳۶	۰, ۲۹	۴, ۹۳	۰, ۲۱	
۷, ۳۵	۶, ۸۳	۶, ۲۷	۰, ۲۲	۶, ۸۲	۰, ۲۲	
۹, ۵۸	۸, ۹	۸, ۲۹	۰, ۲۲	۸, ۹۱	۰, ۲۲	
۰, ۸۵	۰, ۷۹	۰, ۵۶	۰, ۰۷۴	۰, ۷۹	v	

$\hat{\beta}_{11} = 1/68$ برآورد شده، اگر این دو اثر را بطور توازن برای بیماری که ادم ماقولای دارد و طول مدت بیماری آن نیز ۱۵ سال است، در نظر بگیریم یعنی $3/369 = (1/68 \times 15) + (1126 \times 10/0)$ می‌توان چنین استنباط کرد که تغییرات عدم ادم ماقولای به وجود ادم ماقولای و طول مدت بیماری به مدت ۱۵ سال به طور توازن باعث تغییری به اندازه $3/369$ در مقیاس متغیر پنهان می‌شود. با توجه به مقادیر ثابت (عرض از مبدأ) برآورد شده، این میزان تغییر می‌تواند وضعیت شدت رتینوپاتی دیابتی چشم راست یک بیمار را از حالت (رسته) خفیف $6/56 \leq \hat{\theta}_{12} < 4/76$ به حالت (رسته) متوسط و شدید $9/54 \leq \hat{\theta}_{13} < 6/56$ تغییر دهد. پذیرش فرض تقارن در حالت یک و دو متغیره از نظر علم پزشکی نیز درست به نظر نمی‌رسد زیرا پس از ابتلای شخص به بیماری منطقی به نظر نمی‌رسد که احتمال گرایش بیماری به دو سوی مخالف یعنی به سوی بدترین حالت (مرگ یا تخریب) یا به سوی بهترین حالت (بهبودی یا سلامت) یکسان باشد. همچنین اگرستی (۲۰۰۲) بطور کلی و زایری و کاظم نژاد (۲۰۰۶) بطور خاص ارجحیت مدل نامتقارن بر متقارن در چنین موقعی را نشان داده‌اند. استفاده از مدل متقارن نظیر پروبیت و لوجیت به دلیل سادگی برآش، تحلیل وجود برنامه‌های نرم افزاری آماده آسان است و شاید به همین دلیل بیشتر مورد استفاده قرار گرفته‌اند. برتری کاربریت تحلیل بیزی نسبت به تحلیل کلاسیک، قبلّاً نیز اشاره شد و به این دلیل است که در مطالعات مشابه به دلیل همبستگی شدید و نامتقارن بودن داده‌های پاسخ تراکم داده‌ها در یک ربع و روی قطر اصلی جدول بیشتر است، و برای استفاده از تحلیل کلاسیک نیاز به نمونه بیشتری پیدا می‌شود تا پیش فرض‌های کلاسیک صادق باشند، از طرفی در اکثر موقع گردآوری نمونه‌ای به حجم زیاد مستلزم زمان و هزینه بیشتر است. انتخاب غلط توزیع متغیر پنهان ممکن است نتایج گمراه کننده‌ای داشته باشد. در این مقاله برای متغیر پنهان مدل نامتقارن را ارائه کردیم که در صورت نامتقارن بودن پاسخ‌ها بهتر است از این توزیع در مدل‌بندی استفاده شود. علاوه بر آن موقعی که به دلایلی (مثل کمبود بودجه، زمان یا نایاب بودن نمونه) نتوانیم نمونه‌ای به اندازه کافی بزرگ گردآوری کنیم، استفاده از تحلیل کلاسیک جایز نیست و نتایج اشتباہی بدست می‌دهد. به علاوه در این

۱۵۴ تحلیل بیزی مدل دو متغیره تربیتی نامتقارن بر پایه متغیر پنهان

موقعیت بهترین راه حل استفاده از استنباط بیزی است و روش‌هایی مثل ادغام طبقات همیشه محدود نیست (مثلاً مقیاس‌های اسمی)، زیرا معیار خاصی برای ادغام وجود ندارد و در صورت ادغام طبقات نیز دقت تحلیل کم می‌شود.

مراجع

- Agresti, A. (2002), *Categorical Data Analysis*. Wiley, New York 2nd Ed.
- Albert, J. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of American Statistical Association*, **88**, 669-679.
- Bock, R. D. and Jones, L. V. (1968), *The Measurement and Prediction of Judgment and Choice*. Holden Day, San Francisco.
- Biswas, A., and Das, K. (2002), A Bayesian Analysis of Bivariate Ordinal Data: Wisconsin Epidemiologic Study of Diabetic Retinopathy Revisited. *Statistics in Medicine*, **21**, 549-59.
- Chib, S. and Greenberg, E. (1998), Analysis of Multivariate Probit Models. *Biometrika*, **85**, 347-361.
- Gumbel, E. J. (1961), Bivariate Logistic Distribution. *Journal of American Statistical Association*, **56**, 335-349.
- Kim, K. (1995), A Bivariate Cumulative Probit Regression Model for Ordered Categorical Data. *Statistics in Medicine*, **14**, 1341-1352.
- Liang, K. Y. and Zeger, S. L. (1986), Longitudinal Data Analysis using Generalized Linear models. *Biometrika*, **73**, 13-22.

Lipsitz, SR., Kim K., and Zhao, L. (1994), Analysis of Repeated Categorical Data using Generalized Estimating Equation. *Statistics in Medicine*, **13**, 1149-63.

McCullagh, P. (1980), Regression Models for Ordinal Data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109-142.

Nelder, J. A., and Wedderburn, R. W. M. (1972), Generalized Linear Model. *Journal of Royal Statistical Society, Series A*, **135**, 370-84.

OBrien, S. M. and Dunson, D. B. (2004), Bayesian Multivariate Logistic Regression. *Biometrics*, **60**, 739-746.

Rosener, B. (1984), Multivariate Methods in Ophthalmology with Application in other Paired-data Situation. *Biometrics*, **40**, 1025-1035.

Snell, E. J. (1964), A Scaling Procedure for Ordined Categorical Data. *Biometrics*, **20**, 592-607.

Satterthwait, S. P. and Hutchinson, T. P. (1978), A generalization of Gumbel's Bivariate Logistic Distribution. *Metrika*, **25**, 163-170.

Williamson, J. M., Kim, K. and Lipsitz, S. R. (1995), Analyzing Bivariate Ordinal Data using a Global Odds Ratio. *Journal of American Statistical Association*, **90**, 1432-7.

Zayeri, F. and Kazemnejad, A. (2006), A Latent Variable Regression Model for Asymmetric Bivariate Ordered Categorical Data. *Applied Statistics*, **33**, 743-753.