مجله علوم آماری، بهار و تابستان ۱۳۹۱
جلد ۶، شماره ۱، ص ۹۹-۱۱۰

مدل مکانی عام برای پاسخ‌های همبسته پیوسته، ترتیبی و اسمی

سیده فاطمه میری، احسان بهرامی سامانی
گروه آمار، دانشگاه شهید بهشتی

تاریخ دریافت: ۱۳۹۱/۲/۳ تاریخ آخرین بازنگری: ۱۳۹۱/۶/۲۳

چکیده :هدف این مقاله معرفی یک مدل تعمیم‌یافته برای توزیع توام متغیرهای اسمی، ترتیبی و پیوسته با و بدون داده گم‌شده است. فرم‌های بسته‌ای برای تابع درستنمایی مربوط به مدل‌های مکانی عام ارائه می‌شود. همچنین تقریب جو، برای براورد پارامترهای مدل مکانی عام با پاسخ‌های اسمی، پیوسته و ترتیبی با و بدون داده‌های گم‌شده به‌کار برده شده است. برای نشان دادن قابلیت مدل‌های پیشنهادی، مطالعه‌های شبیه‌سازی انجام شده است. همچنین مدل‌های ارائه شده بر روی داده‌های واقعی مربوط به‌آموزش زبان خارجی مورد تحلیل قرار گرفته است.

واژه‌های کلیدی : پاسخ‌های همبسته پیوسته، ترتیبی و اسمی، تابع درستنمایی، داده‌های گم‌شده، متغیر پنهان، مدل مکانی عام.

۱ مقدمه

یکی از موضوع‌های مهمی که مورد توجه محققان علوم آموزشی و علوم پزشکی قرار می‌گیرد، بررسی پیوند بین برداری از متغیرهای پیوسته و گسسته است. یکی از مدل‌هایی که برای این منظور مورد استفاده قرار می‌گیرد، مدل مکانی عام برای داده‌های آمیخته همبسته

آدرس الکترونیک مسئول مقاله: سیده فاطمه میری، sadatmiri88@yahoo.com
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲E۹۹، ۶۲H۱۰

پیوسته و ترتیبی است. در این مدل متغیرهای اسمی یک جدول پیشایندی به‌وسیله تقاطع سطوح‌شان ایجاد می‌کنند که در هر خانه فرض می‌شود متغیرهای پیوسته دارای توزیع نرمال چند متغیره هستند. در واقع مساله اصلی پیوند دو متغیر پیوسته و ترتیبی بر پایه متغیر پنهان و تعیین تاثیر متغیرهای تبیینی روی این متغیرها است. در این مدل یک سطر ماتریس داده‌ها ممکن است برای بعضی یا همه متغیرهای ترتیبی یا پیوسته دارای مقادیر گم‌شده باشند. با توجه به اهمیت مدل مکانی عام اولین بار اولکین وتیت (۱۹۶۱) حالت‌هایی را که تیت (۱۹۵۵) به آن‌ها اشاره کرده بود تعمیم دادند و مدلی تحت عنوان مدل مکانی عام با پاسخ‌های همبسته اسمی وپیوسته مطرح کردند. کاکس (۱۹۷۲) با استفاده از تجزیه توام متغیرهای پیوسته و گسسته به توزیع حاشیه‌ای متغیر پیوسته و توزیع شرطی متغیر گسسته به شرط متغیر پیوسته، مدلی را ارایه نمود که پیوند بین متغیرهای گسسته و پیوسته را مورد بررسی قرار می‌داد. اولین بار رابین (۱۹۷۶) مفهوم گم‌شدگی داده‌ها را معرفی کرد و سپس لیتل و شالتر (۱۹۸۵) براورد ماکسیمم درستنمایی را برای مدل‌های آمیخته پیوسته که دارای پاسخ‌های گم‌شده باشند معرفی کرد. هکمن (۱۹۷۸) مدل همزمان را برای پاسخ‌های آمیخته پیوسته و گسسته ارایه کرد. کاکس و ویرموتس (۱۹۹۲) به مقایسه مدل مکانی عام و مدل کاکس (۱۹۷۲) پرداختند. کرزانوسکی (۱۹۹۳) مدل‌های مکانی را برای داده‌های آمیخته پیوسته وگسسته ارایه کرد. دیگل و همکاران (۱۹۹۴) مدل‌هایی را روی داده‌های طولی آمیخته مطرح کردند. داده‌های طولی آمیخته داده‌هایی هستند که پاسخ‌های آمیخته اسمی و پیوسته برای هر آزمودنی در طول زمان تکرار می‌شوند. فیتز موریس و لیرد (۱۹۹۷) برای حالتی که برخی از پاسخ‌ها دارای گم‌شدگی هستند به‌کار بردند. لیون و کریر (۲۰۰۷) مدل‌های مکانی را برای داده‌های آمیخته پیوسته و ترتیبی با و بدون داده‌های گم‌شده بر اساس مدل‌های همزمان، پرداختند. به‌طور کلی آنچه در این مقاله مورد بررسی قرار گرفته به شرح زیر است: بخش دوم، تعمیم مدل داده‌های آمیخته اسمی، پیوسته و ترتیبی با و بدون داده‌های گم‌شده معرفی شده و سپس براورد پارامترها با روش ماکسیمم درستنمایی به‌دست آمده است. بخش سوم به مطالعات شبیه‌سازی برای نشان دادن قابلیت مدل ارایه شده و در انتها نتایجی که از داده‌های واقعی به‌دست می‌آید ارایه می‌شود. بخش چهارم، یک نتیجه‌گیری کلی از مدل مکانی عام برای پاسخ‌های آمیخته اسمی، پیوسته و ترتیبی بیان می‌شود.

۲ تعمیم مدل داده‌های آمیخته

در این بخش ابتدا مدل مکانی عام را برای حالتی که پاسخ‌های آمیخته پیوسته، ترتیبی و اسمی معرفی شده، ارایه خواهد شد. سپس در ادامه به معرفی تعمیم مدل مکانی عام برای داده‌های آمیخته پیوسته، ترتیبی و اسمی با داده‌های گم‌شده که گم‌شدگی از نوع غیر قابل چشم‌پوشی است بیان می‌شود. فرض کنید $X_1, \dots, X_S$ نشان‌دهنده مجموعه‌ای از متغیرهای اسمی و $Y_1, \dots, Y_c$ نشان دهنده مجموعه‌ای از متغیرهای پیوسته و $Z_1, \dots, Z_Q$ نشان‌دهنده مجموعه‌ای از متغیرهای ترتیبی باشند. اگر هر یک از این متغیرها برای n فرد در نظر گرفته شوند، نتیجه یک ماتریس داده $n \times (S + C + Q)$ بعدی $M = (X, Y, Z)$ است، که‌در آن $X = (X_1, \dots, X_S)$، $Y = (Y_1, \dots, Y_C)$ و $Z = (Z_1, \dots, Z_Q)$ به ترتیب نشان دهنده‌ی بخش‌های اسمی، پیوسته و ترتیبی ماتریس M هستند. توزیع توام $[X, Y, Z]$ به‌صورت

$$[X, Y, Z] = [X][Y|X][Z|X, Y]$$

تجزیه می‌شود، که در آن $[X]$، $[Y|X]$ و $[Z|X, Y]$ به‌ترتیب مشخص کننده توزیع کناری X، توزیع شرطی Y به‌شرط X و توزیع شرطی Z به‌شرط X و Y هستند. همچنین $[Z|X, Y]$ بر مبنای $[Y^*|X, Y]$ توزیع شزطی Y^* به شرط X و Y است.

فرض کنید U بردار $D \times 1$ از متغیرهای اسمی می‌باشد که با dامین مولفه از آن دارای s_d حالت ممکن $(d = 1, \dots, D)$ است. بردار U یک جدول پیشایندی با $S = \prod_{d=1}^{D} s_d$ حالت ممکن برای مقادیر U تعریف می‌کند. یک بردار $S \times 1$، $x = (X_1, \dots, X_s)$ می‌تواد تعریف شود، که اگر U در حالت s قرار گیرد X_s برابر یک و در غیر این صورت برابر صفر است. در این‌صورت هر سطر U فقط شامل یک ۱ است. تحت مدل مکانی عام متغیر X دارای توزیع چند جمله‌ای است، یعنی $X|\pi \propto \prod_{s=1}^{S} \pi_s^{x_s}$ که در آن $\pi = (\pi_1, \dots, \pi_S)^T$ یک بردار از احتمال‌های خانه‌های مربوط به جدول پیشایندی است. برای $[y, y^*|x_{(s)}]$، که y برداری $C \times 1$ و y^* برداری $Q \times 1$، یک مدل مکانی تعمیم یافته با میانگین‌های $E(Y|x_s) = \mu_s$ و $E(Y^*|X_s) = \mu_s^*$ و ماتریس‌های کواریانس $Cov(Y^*|x_s) = Cov(Y^*) = \Sigma^*$، $Cov(Y|x_s) = Cov(Y) = \Sigma$ و $Cov(Y, Y^*|x_s) = Cov(Y, Y^*) = \Sigma_{Y,Y^*}$ هستند. ارتباط بین متغیر پنهان و متغیر ترتیبی به‌وسیله مدل آستانه‌ای تعریف می‌شود، یعنی اگر $\alpha_q^{\ell_q - 1} < Y_q^* < \alpha_q^{\ell_q}$ آنگاه $Z_q = a_q^{\ell_q}$، که در آن $\alpha_q^{\ell_q}$، $q = 1, \dots, Q$ نقاط آستانه‌ای، $\alpha_q^0 = -\infty$، $\alpha_q^{\ell_q + 1} = +\infty$ و $\{\alpha_q^1, \cdots, \alpha_q^{\ell_q}\}$ نقاط نامعلوم هستند. یک رابطه ترتیبی برای نقاط α به‌صورت $a_q^1 < \cdots < a_q^{L_q + 1}$ است و

تحت مدل مکانی عام توزیع Y^* به شرط Y و x_s نرمال چند متغیرهٔ

$$Y^*|Y, x_s \sim N(\mu_s^* + \Sigma_{y,y^*}^T \Sigma^{-1}(y - \mu_s), \Sigma^* - \Sigma_{y,y^*}^T \Sigma^{-1} \Sigma_{y,y^*})$$

است. ماتریس کواریانس به فرم $DRD = \Sigma^* - \Sigma_{y,y^*}^T \Sigma^{-1} \Sigma_{y,y^*}$، است، که در آن D ماتریس قطری از انحراف معیار شرطی d_q و R ماتریس متقارن $Q \times Q$ از همبستگی چند رشته‌ای Y^* به‌شرط X_s و Y است. برای کاهش دادن تعداد پارامترهای مدل، S ثابت در نظر گرفته شده است، که در آن

$$\mu_s = \xi + \xi_s \ \ , \ \ \mu_s^* = \xi^* + \xi_s^*, \quad s = 1, \ldots, S-1$$

همچنین $\mu_S = \xi$ و $\mu_S^* = \xi^*$ به‌ترتیب مبانگین‌های Y و Y^* برای حالت s هستند. استاندارد $[Y^*|x_s, Y]$ برای $s \neq S$ به‌صورت

$$D^{-1}(y^* - \mu_s^* - \Sigma_{yy^*}^T \Sigma^{-1}(y - \mu_s)) = D^{-1}(y^* - \xi^*) - \tau_s - B(y - \xi)) \qquad (1)$$

انجام می‌شود، که در آن

$$\tau_s = D^{-1}\xi_s^* - B\xi_s \ , \ \ B = D^{-1}\Sigma_{yy^*}^T \Sigma^{-1}$$

عبارت (۱) دارای توزیع نرمال چند متغیره با میانگین صفر و ماتریس کواریانس R است. به‌طور مشابه استاندارد نقاط آستانه‌ای شرطی $\{\alpha_q^1, \cdots, \alpha_q^{L_q}\}$، $(q = 1, \ldots, Q)$ به‌صورت $\gamma_q^{\ell_q} - \tau_{sq} - \beta_q^T y$ که در آن $(\frac{\xi_q^*}{d_q} - \beta_q^T \xi) - \frac{\alpha_q^{\ell_q}}{d_q} = \gamma_q^{\ell_q}$، $\tau_{sq} = \frac{\xi_{sq}^*}{d_q} - \beta_q \xi_{sq}$، q امین عنصر $\beta_q = d_q^{-1} S_{y,y_q^*} \Sigma^{-1}$ $\beta_q^T = (\beta_{q1}, \cdots, \beta_{qC})$، qامین ردیف B است. توجه شود $\tau_{Sq} = 0$، $\gamma_q^0 = -\infty$ و $\gamma_q^{L_q+1} = +\infty$ هستند. واضح است که

$$P(Z = \ell | x = x_s, y, \theta) = \int_S \phi_Q(v|R) dv,$$

که در آن $\phi_Q(\cdot|R)$ تابع چگالی توزیع نرمال با میانگین صفر و ماتریس کواریانس R، $\ell = (\ell_1, \ldots, \ell_Q)^T$، $S = \{(v_1, \ldots, v_Q) : \nu_{sq}^{\ell_q - 1} < v_q < \nu_{sq}^{\ell_q}, q = 1, \ldots, Q\}$، که $\nu_{sq}^{\ell_q} = \gamma_q^{\ell_q} - \beta_q^T y - \tau_{sq}$ و $\theta = (\mu, \Sigma, R)$ هستند. چگالی توام $[x, y, z]$ به‌صورت

$$P(x = x_s, y, z = \ell | \theta) = \pi_s \phi(y - \mu_s | \Sigma) \int_S \phi_Q(v|R) dv$$

تعریف می‌شود و متغیرهای پیوسته Y^*، $I_{R_y^*}$ و $I_{R_z^*}$ به‌ترتیب نشان‌دهنده متغیر پنهان برای پاسخ ترتیبی Z_i، متغیر پنهان مربوط به مکانیسم گم‌شدگی Y_i و متغیر پنهان مربوط به

مکانیسم گم‌شدگی Z_i هستند. از طرفی دیگر متغیرهای نشانگر $I_{R_{y_i}}$ و $I_{R_{z_i}}$ به‌ترتیب به‌صورت

$$I_{R_{y_i}} = \begin{cases} 1 & R^*_{y_i} > 0 \\ 0 & o.w, \end{cases} \qquad I_{R_{z_i}} = \begin{cases} 1 & R^*_{z_i} > 0 \\ 0 & o.w. \end{cases}$$

تعریف می‌شوند، که در آن‌ها $R^*_{y_i}$ و $R^*_{z_i}$، متغیرهای مربوط به R_{y_i} و R_{z_i} هستند. یک بودن تابع نشانگر نشان‌دهنده این است که متغیر پاسخ مشاهده شده است و صفر بودن آن نشان دهنده این است که متغیر پاسخ گم‌شده است. مدل مکانی عام مربوط به پاسخ‌های پیوسته و ترتیبی که دارای گم‌شدگی هستند به‌صورت

$$Y^*_i = \beta'_1 T_{i1} + \varepsilon_{i1}, \qquad Y_i = \beta'_2 T_{i2} + \varepsilon_{i2}$$
$$R^*_{Y_i} = \alpha'_1 T_{i3} + \varepsilon_{i3}, \qquad R^*_{Z_i} = \alpha'_2 T_{i4} + \varepsilon_{i4}$$

در نظر گرفته می‌شوند، که در آن $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$ دارای توزیع نرمال چند متغیره با میانگین صفر و ماتریس کواریانس Σ_{1234} است. بنابراین پارامترهای مدل در حالتی که بردار متغیرهای Z و Y عبارتند از β'_1، β'_2، η'_1 و η'_2 و نقاط آستانه‌ای $\alpha^1_q < \ldots < \alpha^{L_q}_q$ و ضریب همبستگی R_{1234} هستند. به خاطر نشان‌پذیر بودن پارامترهای مدل Σ_{1234}، $\Sigma_{y^*,y^*} = Cov(y^*, y^*) = I$، $\Sigma_{R^*_z,R^*_z} = Cov(R^*_z, R^*_z) = I$ و $\Sigma_{R^*_y,R^*_y} = Cov(R^*_y, R^*_y) = I$ درنظر گرفته می‌شوند و $\Sigma_{y^*,R^*_z} = Cov(y^*, R^*_z)$، $\Sigma_{y^*,R^*_y} = Cov(y^*, R^*_y)$ و در نهایت $\Sigma_{R^*_z,R^*_y} = Cov(R^*_z, R^*_y)$، $\Sigma_{y,R^*_z} = Cov(y, R^*_z)$. لازم به ذکر است پارامترها در صورت صفر بودن ماتریس‌های Σ_{y^*,R^*_z}، Σ_{y^*,R^*_y}، $\Sigma_{R^*_z,R^*_y}$ و Σ_{y,R^*_z} مکانیسم گم‌شدگی، از نوع قابل چشم‌پوشی است.

۱.۲ تابع درستنمایی مدل مکانی عام برای داده‌های کامل

فرض کنید نمونه تصادفی به اندازه N از مدل داده آمیخته تعمیم یافته[1] (GMDM) مشاهده شده است. لگاریتم تابع درستنمایی داده‌ها به‌صورت

$$\log L(\pi_s, \mu, \Sigma, R | x_s, y, z) = \ell(\pi_s | x_s) + \ell(\mu, \Sigma | x_s, y) + \ell(R | x_s, y, z)$$

[1] General Mixed Data Model

به‌طوری‌که

$$\ell(\pi|X) = \sum_{s=1}^{S} n_s \log(\pi_s)$$

$$\ell(\mu, \sigma|X, Y) = \frac{-N}{2} \log|\Sigma| - \frac{1}{2} \sum_{s=1}^{S} \sum_{i(s) \in N} (y_{i(s)} - \mu_s)^T \Sigma^{-1} (y_i - \mu_s)$$

$$\ell(R|X, Y, Z) = \sum_{i=1}^{N} \sum_{s=1}^{S} \log \sum_{\epsilon_q = 0}^{1} (-1)^{\sum_{q=1}^{Q} \epsilon_q + Q} \Phi_Q(\ldots, v_{i(s,\ell)}^{\ell_q - \epsilon_q}, \ldots |R)$$

که در آن $\Phi_Q(\ldots, v_{i(s,\ell)}^{\ell_q-\epsilon_q}, \ldots|R)$ تابع توزیع نرمال چند متغیره با میانگین صفر و ماتریس همبستگی Rاست.

۲.۲ تابع درستنمایی مدل مکانی عام برای داده‌های گم‌شده

برای بررسی و تحلیل پاسخ‌های همبسته برداری از متغیرهای پیوسته و برداری از متغیرهای ترتیبی بر پایه متغیر پنهان لازم است تابع درستنمایی پاسخ‌ها تعیین شود که محاسبه آن در شرایط خاص با پیچیدگی و دشواری همراه است. برای حل این مساله، می‌توان تابع درستنمایی را تقریب زد (جو، ۱۹۹۵). برای این منظور فرض کنید $X = (X_1, \ldots, X_m)$ برای $m \geq 3$، دارای توزیع نرمال چند متغیره با بردار میانگین صفر، واریانس‌های ۱ و ماتریس همبستگی Ω باشد.

فرض کنید $I_i = I(\omega_i \geq X_i \geq x_i)$، برای $i = 1, \ldots, m$ یک تابع نشانگر باشد به‌طوری‌که $E(i) = \Phi(x_i) - \Phi(\omega_i)$، که در آن Φ تابع توریع نرمال استاندارد است. اکنون با توجه به تجزیهٔ

$$\begin{aligned} P(\omega_1 \geq X_1 \geq X_2, \ldots, \omega \geq X_m \geq x_m) &= P(\omega_1 \geq X_1 \geq x_1, \omega_2 \geq X_2 \geq x_2) \\ &\times \Pi_{k=3}^{m} P(\omega_k \geq X_k \geq x_k | \omega_1 \geq X_1 \geq x_1, \ldots, \omega_{k-1} \geq X_{k-1} \geq x_{k-1}) \quad (2) \end{aligned}$$

در مرحله اول، تقریب عبارت دوم رابطه (۲) به‌صورت

$$E(I_k | I_1 = 1, \ldots, I_{k-1} = 1) = E(I_k) + \Omega_{21} \Omega_{11}^{-1} (1 - E(I_1), \ldots, 1 - E(I_{k-1}))^T$$

است، که در آن Ω_{21} یک بردار سطری شامل اعضای $Cov(I_i, I_k)$، $i = 1, \ldots, k-1$ و Ω_{11} یک ماتریس $(k-1) \times (k-1)$ با درایه‌های

$Cov(I_i, I_j) = E(I_i, I_j) - E(I_i)E(I_j)$ و $1 \leq j \leq k-1$ و $i \geq 1$ است، لازم به ذکر است که $E(I_i, I_j) = P(\omega_i \geq X_i \geq x_i, \omega_j \geq X_j \geq x_j)$. برای اطلاعات بیشتر به جوی (۱۹۹۵) رجوع شود.

تابع درستنمایی برای پاسخ‌های آمیخته پیوسته و ترتیبی به‌طوری که پاسخ‌ها، دارای گم‌شدگی هستند و متغیر اسمی همواره مشاهده شده باشند، برای افراد گوناگون وابسته به الگوی گم‌شدنشان به‌گونهٔ متفاوتی بایستی به‌دست آید. با در نظر گرفتن یک متغیر پاسخ Y، Z و X تابع درستنمایی در چهار حالت به‌دست آورده‌شده‌است که دارای ماتریس کواریانس باپارامترهای σ (واریانس متغیر پاسخ Y)، ρ_{12} (ضریب همبستگی بین Y و Y^*)، ρ_{13} (ضریب همبستگی بین Y و R^*_Y)، ρ_{14} (ضریب همبستگی بین Y و $R^*_{Y^*}$)، ρ_{23} (ضریب همبستگی بین Y^* و R^*_Y)، ρ_{24} (ضریب همبستگی بین Y^* و $R^*_{Y^*}$) و ρ_{34} (ضریب همبستگی بین R^*_Y و $R^*_{Y^*}$) است. پارامترهای η_1، η_2، β'_1 و β'_2 ضرایب رگرسیونی مدل مورد نظر هستند و همچنین نقاط آستانه‌ای مربوط به متغیر تبیینی X برآورد می‌باشد. در این‌جا فقط پاسخ Y برای فرد iام مشاهده شده مورد بررسی قرار می‌گیرد و می‌توان آن‌را برای حالت‌های دیگر (۱- فقط پاسخ Z برای فرد iام مشاهده شده است. ۲- هر دو پاسخ Y و Z برای فرد iام مشاهده نشده باشند. ۳- هردو پاسخ Y و Z برای فرد iام مشاهده شده باشند.) تعمیم داد.

$$\begin{aligned}
L_i &= f(y_i, I_{R_{Y_i}} = 1, I_{R_{z_i}} = 0) \\
&= f(y_i | X_i = x_i) P(X_i = x_i) [P(I_{R_{z_i}} = 0 | y_i, X_i = x_i) \\
&- P(I_{R_{z_i}} = 0, I_{R_{y_i}} = 0 | y_i, X_i = x_i)] \\
&= f(y_i | X_i = x_i) P(X_i = x_i) [P(R^*_{z_i} < 0 | y_i, X_i = x_i) \\
&- P(R^*_{z_i} < 0, R^*_{y_i} < 0 | y_i, X_i = x_i)] \\
&= f(y_i | X_i = x_i) P(X_i = x_i) \\
&\times [\Phi(\frac{-\alpha'_2 T_{i4} - \frac{\rho_{24}}{\sigma}(y_i - \beta'_2 T_{i2})}{\sqrt{1 - \rho^2_{24}}}) - \Phi_{12}(\frac{-\alpha'_2 T_{i4} - \frac{\rho_{24}}{\sigma}(y_i - \beta'_2 T_{i2})}{\sqrt{1 - \rho^2_{24}}}, \\
&\quad \frac{-\alpha'_1 T_{i3} - \frac{\rho_{23}}{\sigma}(y_i - \beta'_2 T_{i2})}{\sqrt{1 - \rho^2_{23}}}; \frac{\rho_{34} - \rho_{24}\rho_{23}}{\sqrt{1 - \rho^2_{23}}\sqrt{1 - \rho^2_{24}}})]
\end{aligned}$$

که در آن $\Phi_{12}(\cdot, \cdot; \cdot)$ تابع توزیع تجمعی نرمال متغیرهای پنهان مربوط به مکانیزم گم‌شدگی Y_i و مکانیزم گم‌شدگی Z_i است. تابع درستنمایی که از ضرب تمام درستنمایی‌ها به‌دست می‌آید را می‌توان در نرم‌افزار R، با دستور "$nlminb$" بهینه کرد.

۳ مطالعه شبیه‌سازی

هدف از این شبیه‌سازی، بررسی مدل مکانی عام با داده های گم‌شده است. با فرض آنکه $C = Q = L = ۱$ و $s = ۲$ باشند، Y و Y^* (متغیر پنهان) دو متغیر پیوسته، Z یک متغیر ترتیبی با دو سطح و X یک بردار اسمی دوتایی در نظر گرفته شده‌اند. در این مدل فرض می‌شود متغیر ترتیبی دارای گم‌شدگی است. پارامترهای مدل عبارتند از $\theta = (\pi, \mu^T, \sigma^۲, \gamma, \beta, \tau)$، که در آن، $\mu^T = (\mu_۱, \mu_۲)$ امید Y به‌شرط $x = x_{(s)}$ به‌ازای $s = ۱, ۲$، γ استاندارد $R^*_{z_i}$،

$$\gamma = \frac{R^*{}_{z_i}}{\sqrt{۱ - \rho^۲}} - (\frac{\mu^*_۲}{\sqrt{۱ - \rho^۲}} - \beta\mu_۲), \quad \beta = \frac{\rho}{\sigma\sqrt{۱ - \rho^۲}}, \quad \tau = \frac{\xi^*}{\sqrt{۱ - \rho^۲}} - \beta\xi$$

و ρ ضریب همبستگی بین Y و Y^* است. در این شبیه‌سازی نمونه‌هایی به حجم ۲۵۰، ۵۰۰ و ۱۰۰۰ با ۲۰۰۰ مرتبه تکرار از GMDM با مقادیر واقعی $\pi = ۰/۵$، $\mu_۱ = \mu^*_۱ = ۰$، $\mu_۲ = \mu^*_۲ = ۰/۵$، $\sigma^۲ = ۱$، $\rho = ۰/۵$، $\gamma = -۰/۲۸۹$، $\beta = ۰/۵۷۷$ و $\tau = -۰/۲۸۹$ تولید شده است. با توجه به نتایج ارایه شده در جدول ۱ ملاحظه می‌شود، هر اندازه حجم نمونه زیاد باشد براورد پارامترها به مقدار واقعی نزدیک‌تر و اریبی کم‌تر است. همچنین با افزایش حجم نمونه کارایی نسبی افزایش می‌یابد و مقادیری که بزرگتر از یک هستند نشان می‌دهد مقدار براورد پارامتر کاراتر از مقدار میانگین پارامترها است.

جدول ۱: برآوردهای ماکسیمم درستنمایی پارامترها با ۲۰۰۰ مرتبه تکرار از GMDM با داده‌های گم‌شده

پارامتر	مقدار پارامتر	حجم نمونه	میانگین براورد	اریبی	کارایی
γ	۰/۸۶۶	۲۵۰	۰/۸۷۱۶	−۱/۶۸۲۶	۰/۹۵۸۸
		۵۰۰	۰/۸۴۳۲	−۰/۸۸۶۴	۱/۰۱۴۷
		۱۰۰۰	−۰/۸۹۰۰	−۰/۶۴۶۰	۱/۲۰۰
β	۰/۵۷۷	۲۵۰	۰/۵۹۰۰	−۲/۲۵۳۰	۰/۹۹۴۷
		۵۰۰	۰/۵۸۷۱	−۱/۷۵۰۱	۱/۰۰۰۱
		۱۰۰۰	۰/۵۸۹۲	−۲/۱۱۴۰	۱/۰۱۶۶
τ	−۰/۲۸۹	۲۵۰	−۰/۲۹۴۵	−۱/۹۰۳۰	۰/۹۲۲۰
		۵۰۰	−۰/۲۹۳۲	−۱/۴۵۳۰	۰/۹۴۵۵
		۱۰۰۰	−۰/۲۹۰۰	۰/۳۴۶۰	۱/۰۰۶۵

جدول ۲: براورد پارامترها و خطای استاندارد مدل کامل GMDM

	پارامتر	براورد	انحراف معیار
$SEX \times LAN$			
فرانسوی و مرد	π_1	۰/۱۴۳	۰/۰۲۳
اسپانیایی و مرد	π_2	۰/۱۵۲	۰/۰۲۴
آلمانی و مرد	π_3	۰/۲۵۱	۰/۰۲۹
فرانسوی و زن	π_4	۰/۱۳۰	۰/۰۲۲
اسپانیایی و زن	π_5	۰/۱۲۶	۰/۰۲۲
آلمانی و زن	π_6	۰/۱۹۹	۰/۰۲۶
فرانسوی و مرد	μ_{11}	۸۴/۹۳۹	۲/۳۰۰
	μ_{12}	۲/۸۰۳	۰/۱۰۹
اسپانیایی و مرد	μ_{21}	۷۸/۵۴۳	۲/۲۳۴
	μ_{22}	۲/۶۱۰	۰/۱۰۶
آلمانی و مرد	μ_{31}	۸۲/۱۷۲	۱/۷۳۵
	μ_{32}	۲/۸۰۰	۰/۰۸۲
فرانسوی و زن	μ_{41}	۸۲/۷۶۷	۲/۴۱۳
	μ_{42}	۲/۷۶۹	۰/۱۱۴
اسپانیایی و زن	μ_{51}	۸۴/۲۴۱	۲/۴۵۴
	μ_{52}	۲/۸۹۱	۰/۱۱۶
آلمانی و زن	μ_{61}	۸۵/۵۰۰	۱/۹۴۸
	μ_{62}	۲/۷۴۳	۰/۰۹۲
واریانس	σ_1^2	۱۷۴/۶۱۳	۱۶/۲۴۷
	σ_2^2	۰/۳۹۲	۰/۰۳۶
ضریب همبستگی	ρ	۰/۰۵۹	۰/۹۶
نقاط برش	γ^1	۰-/۵۹۷	۰/۲۳
	γ^2	۰/۴۰۰	۰/۲۲۱
اثر رگرسیونی FLAS	β_1	۰-/۶۸۰	۰/۱۱۳
اثر رگرسیونی HGPA	β_2	۰-/۰۱۲	۰/۰۰۴
	τ_1	−۰/۸۵۸	۰/۲۷۳
	τ_2	−۰/۵۸۴	۰/۲۴۶
اثر $SEX \times LAN$	τ_3	−۰/۵۷۸	۰/۲۹
	τ_4	−۰/۲۶۲	۰/۲۹۷
	τ_5	۰/۴۰۳	۰/۲۷۹

۱.۳ مثال کاربردی برای داده‌های کامل

این بخش شامل کاربرد روان‌سنجی زبان خارجی که از مطالعه شفر (۱۹۹۷)، به‌دست آمده ارائه می‌شود. هدف اصلی از مطالعه زبان خارجی، بررسی ویژگی روان‌سنجی[۲] (FLAS) که توسط ریموند و رابرتز (۱۹۸۳) برای پیشگویی مهارت در FLAS گزارش یافته است مورد بررسی قرار می‌گیرد. آن‌ها ۲۳۱ دانشجو که در دوره‌های زبان خارجی در دانشگاه پنسیلونیا در سال ۱۹۸۰ با پر کردن فرم‌هایی که شامل متغیرهای زیر هستند را در نظر گرفتند. ۱ – LAN، یک متغیر اسمی با سه سطح، متناظر با رشته‌های تحصیلی دانشجویان (۱– زبان فرانسوی ۲– زبان اسپانیایی ۳– زبان آلمانی). ۲– SEX، یک متغیر اسمی با دوسطح، متناظر با جنسیت افراد (۱– مذکر ۲– مونث). ۳– FLAS، یک متغیر پاسخ پیوسته، نمرات دانشجویان. ۴– HGPA، یک متغیر پاسخ پیوسته، معدل دانشجویان در دوره دبیرستان. ۵– GRD[۳]، یک متغیر پاسخ ترتیبی با سه سطح، نمرات دانشجویان زبان (۱– نمره C یا زیر C ۲– نمره B ۳– نمره A). شفر (۱۹۹۷) به تحلیل یک مجموعه بزرگ داده‌ها که شامل اطلاعات بر روی یک تعداد متغیرهای دیگر همانند نمرات آزمون استعداد مدرسه بود پرداخته است. هدف او از انجام این کار، به‌کار بردن مدل مکانی عام بوده است. حال هدف از مطالعه، بررسی ارتباط بین متغیرهای آمیخته با مدل‌بندی کردن بر روی توزیع توام، بر اساس GMDM است. با توجه به این که GRD نتیجه مورد انتظار است، متغیرها را با روش GMDM در حالتی که گم‌شدگی وجود ندارد به‌صورت

$$[GRD, FLAS, HGPA, LAN, SEX] = [GRD|FLAS, HGPA, LAN, SEX] \\ \times[FLAS, HGPA|LAN, SEX] \times [LAN, SEX]$$

گروه‌بندی می‌شود. نتیجه برازش مدل GMDM کامل به‌وسیله‌ی براورد ماکسیمم درستنمایی در جدول ۲ نشان داده شده است. LAN و SEX به‌تریب سه سطحی و دو سطحی هستند، تعداد خانه‌های جدول پیشایندی $S = 6$ است. با بررسی براورد احتمال‌ها مشاهده می‌شود کم‌ترین براورد احتمال ۰/۱۲، در رشته زبان اسپانیایی و جنس مونث و بیشترین آن، ۰/۲۵ در رشته زبان آلمانی و جنس مذکر است. بررسی براورد میانگین‌ها، در متغیر پیوسته FLAS، نشان می‌دهد دانش‌آموزان مذکر که زبان اسپانیایی را مطالعه می‌کنند دارای کمترین نمره و دانش آموزان مونث که زبان آلمانی را مطالعه می‌کنند دارای بیشترین نمره هستند. ضریب همبستگی بین دو متغیر پیوسته FLAS و HGPA، ($\hat{\rho} = 0/0591$) یک ارتباط ضعیف بین

[۲] Foreign Language Attitude Scale

[۳] Grad in The Foreign Language

دو متغیر را نشان می‌دهد. با توجه به جدول ۲ آماره Z برای حالت رگرسیونی FLAS برابر ۲/۸۲۸− و HGPA برابر ۶/۰۲− هستند. این نشان می‌دهد، متغیرهای پیوسته، توانایی پیشگویی متغیر ترتیبی (GRD) را دارند.

۴ بحث و نتیجه‌گیری

یکی از مدل‌های مهم در پیوند بین برداری متغیرهای پیوسته و گسسته، مدل مکانی عام برای داده‌های آمیخته همبسته پیوسته و ترتیبی است. یکی از مسائل به‌کارگیری این مدل، وجود پاسخ‌های گم‌شده و تشخیص مکانیسم گم‌شدگی پاسخ‌ها و عوامل موثر در گم‌شدن آن‌ها است. دراین مقاله این مدل برای پاسخ‌های آمیخته پیوسته، ترتیبی و اسمی با و بدون داده‌های گم‌شده بررسی شد. در این مدل متغیرهای اسمی، یک جدول پیشایندی به‌وسیله تقاطع سطوح‌شان تشکیل می‌دهند و متغیرهای پیوسته و پنهان متناظر با این متغیرها درون خانه‌های جدول قرار می‌گیرند. مدل مکانی عام برای پاسخ‌های آمیخته پیوسته، ترتیبی و اسمی با و بدون داده‌های گم‌شده مواردی هستند که در آینده می‌توانند مورد کاوش قرار گیرند.

مراجع

Bahrami Samani, E., Ganjali, M. (2008), A Latent Variable Model for Mixed Continuous and Ordinal Responses, Journal of Statistical Theory and Applications,*Journal of Statistical*, **7**, 337-348.

Cox, D. R. (1972), The Analysis of Multivariate Binary Data, *Applied Statistics*, **21**, 113-126.

Cox, D. R, and Wermuth, N. (1992), Response Models for Mixed Binory and Quantitave Variables, *Biometrika*, **79**, 441-461.

Diggle, P. J. and Kenward, M. G. (1994), Informative Dropout in Longitudinal Data Analysis. *Journal of Applied Statistics*, **43**, 49-93.

Fitzmaurice, G. M. and Laird, N. M, (1995), Regresion Models for Bivariate Discrete and Continuous Outcome with Clustering, *Journal of the American Statistical Association*, **90**, 845-852.

Fitzmaurice, G. M. And Laird, N. M, (1997), Regression Models for Mixed Discrete and Continuous Responses with Potentially Missing Values. *Biometrika*, **53**, 110-122.

Heckman, J. J. (1978), Dummy Endogenous Variables in a Simultaneous Equation System, *Biometrika*, **88**, 551-561.

Joe, H. (1995), Approximations Multivariate Normal Rectangle Probabilites Based on Conditional Expectation. *Journal of the American Statistical Association*, **90**, 957-967.

Krzanowski, W. J. (1993), The Location Model for Mixtures of Categorical and Continuous Variables. *Journal of Classification*, **10**, 25-49.

Leon. A. R. and Carrer. K. C. (2007), General Mixed-Data Model: Extension of General Location and Grouped Continuous Models, *The Canadian Journal of Statistics*, **35**, 533-548.

Little, R. J. A. and Schluchter, M. D. (1985), Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values, *Biometrika*, **72**, 492-512.

Olkin, I. and Tate, R. F. (1961), Multivariate Correlation Models with Mixed Discrete and Continuous Variables, *The Annals of Mathematical Statistics*, **32**, 448-465; Correction. *The Annals of Mathematical Statistics*, **36**, 343-344.

Raymond, M. R. and Roberts, D. M. (1983), Development and Validation of a Foreign Language Attitude Scale, *Educational and Psychological Measurement*, **43**, 1239-1246.

Rubin, D. B. (1976), Inference and Missing Data. *Biometrika*, **82**, 669-710.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman, Hall, CRC, Boca Raton, Florida.

Tate, F. R. (1955), The Teory of Correlation Between two Continuous Variabls when one is Dichatomized, *Biometrika*, **42**, 205-216.