مجله علوم آماری، بهار و تابستان ۱۳۹۲

جلد ۷، شماره ۱، ص ۷۷-۱۰۲

مدل رگرسیونی بیزی با متغیر پاسخ پواسون آمیخته متناهی دومتغیره

افشین فلاح، مهسا نادی‌فر، رامین کاظمی

گروه آمار، دانشگاه بین المللی امام خمینی

تاریخ دریافت: ۱۳۹۱/۱۰/۳ تاریخ آخرین بازنگری: ۱۳۹۲/۴/۳۰

چکیده : در این مقاله تحلیل رگرسیونی با متغیر پاسخ دارای توزیع پواسون دومتغیره آمیخته با رهیافت بیزی مورد بررسی قرار گرفته است. نشان داده شده است که به دلیل شکل پیچیده تابع درستنمایی مبتنی بر توزیع پواسون دومتغیره، توزیع پسین فاقد شکل بسته بوده و پیچیده است. از این رو، توزیع‌های پسین شرطی کامل پارامترها محاسبه و الگوریتم گیبز برای نمونه‌گیری از توزیع پسین ارائه شده است. به‌منظور ارزیابی مدل بیزی پیشنهادی، مطالعه‌ای شبیه‌سازی انجام شده و کارایی براوردگرهای بیزی پیشنهادی برای پارامترهای مدل با همتای بسامدی آن‌ها مقایسه شده است. همچنین، نحوه کاربست رهیافت بیزی پیشنهادی در قالب یک مثال کاربردی شرح داده شده و کارایی آن مورد ارزیابی قرار گرفته است.

واژه‌های کلیدی : رگرسیون پواسون، توزیع آمیخته، الگوریتم EM، الگوریتم گیبز.

---

* آدرس الکترونیک مسئول مقاله: افشین فلاح، a.fallah@sci.ikiu.ac.ir

* کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲J۱۰۵ و ۱۴M۹۹

۱ مقدمه

مدل رگرسیون پواسون از تکنیک‌های بسیار پرکاربرد در زمینه تحلیل داده‌های شمارشی است. این مدل نمونه‌ای از مدل‌های خطی تعمیم‌یافته است، که در آن متغیر وابسته شمارشی بوده و از توزیع پواسون پیروی می‌کند. معادله رگرسیون پواسون، یک متغیر شمارشی با نرخ وقوع خاص را به مجموعه‌ای از متغیرهای تبیینی[۱] مربوط می‌سازد. برابری واریانس متغیر وابسته با میانگین آن یکی از مفروضات اصلی تحلیل رگرسیون پواسون است. اما در بسیاری از موارد مشاهدات پاسخ بیش‌پراکنده‌اند، به این معنی که واریانس مشاهدات به طور معنی‌داری از میانگین آن‌ها بزرگ‌تر است. در این شرایط برازش مدل رگرسیونی پواسون به داده‌ها مناسب نیست. در این مواقع می‌توان از تحلیل رگرسیون دوجمله‌ای منفی کمک گرفت (گاردنر و همکاران، ۱۹۹۵). وجود صفرهای بیش از حد در مشاهدات متغیر پاسخ نیز یکی دیگر از مشکلات در تحلیل داده‌های شمارشی است. لمبرت (۱۹۹۲) مدل رگرسیون پواسون در صفر متورم[۲] را معرفی کرد. مدل رگرسیون دوجمله‌ای منفی در صفر متورم نیز توسط وانگ (۲۰۰۳) و گارای و همکاران (۲۰۱۱) مورد مطالعه قرار گرفته‌است. یکی از راه‌های مقابله با مشکل وجود صفرهای بیش از حد در مشاهدات پاسخ، استفاده از توزیع پواسون دومتغیره است. مدل‌های در صفر متورم در حالت دومتغیره توسط کارلیس و انسوفراس (۲۰۰۳) و گورمو و والدر (۲۰۰۸) مورد مطالعه قرار گرفته‌اند. اخیرا برمودز (۲۰۰۹) و برمودز و کارلیس (۲۰۱۱) درباره نحوه استفاده از مدل‌های رگرسیون در صفر متورم دو و چندمتغیره برای رتبه‌بندی بیمه اتومبیل مطالعاتی انجام داده‌اند. آن‌ها مدل رگرسیون پواسون دومتغیره را برای بررسی همبستگی ذاتی دو ادعا در صنعت بیمه مورد استفاده قرار دادند. یکی از مزیت‌های استفاده از توزیع‌های دومتغیره این است که می‌توان همبستگی بین دو متغیر پاسخ مختلف را در نظر گرفت و به نتایج مطلوب‌تری دست یافت. گرچه این مدل‌ها برای داده‌های در صفر متورم مناسب هستند، اما برای مدل‌بندی مشاهداتی که بیش‌پراکنده نیز باشند، از کارایی مطلوبی

[۱] Explanatory

[۲] Zero-inflated

برخوردار نیستند. از این رو، برمودز و کارلیس (۲۰۱۲) مدل رگرسیون آمیخته متناهی پواسون دومتغیره را مطرح و نحوه براورد پارامترهای مدل را بر اساس رهیافت بسامدی ماکسیمم درستنمایی و به کمک الگوریتم EM شرح دادند.

در این مقاله تحلیل رگرسیونی با متغیر پاسخ دارای توزیع پواسون آمیخته دومتغیره با رهیافت بیزی مورد توجه قرار گرفته است. برای این منظور، با در نظر گرفتن توزیع‌های پیشین مناسب برای پارامترها و بر اساس تابع درستنمایی داده‌های کامل، توزیع پسین پارامترهای مدل محاسبه شده است. نشان داده شده است که به‌دلیل شکل پیچیده تابع درستنمایی مبتنی بر توزیع پواسون دومتغیره، توزیع پسین فاقد شکل بسته بوده و پیچیده است. از این رو، توزیع‌های پسین شرطی کامل پارامترها محاسبه و الگوریتم گیبز برای نمونه‌گیری از توزیع پسین استفاده شده است. همچنین، کارایی مدل بیزی پیشنهادی در براورد پارامترهای مدل در مقایسه با مدل بسامدی رقیب مورد ارزیابی قرار گرفته است.

در بخش ۲ مدل رگرسیون آمیخته متناهی پواسون دومتغیره و داده‌های ناقص و کامل به طور اجمالی معرفی و سپس رهیافت بسامدی براورد پارامترهای مدل بر اساس روش ماکسیمم درستنمایی و الگوریتم EM، شرح داده شده است. در بخش ۳ رهیافت بیزی تحلیل رگرسیون آمیخته متناهی پواسون دومتغیره مورد بحث قرار گرفته و نحوه تعیین توزیع پیشین، محاسبه توزیع پسین، نمونه‌گیری از توزیع پسین و غیره شرح داده شده است. در بخش ۴، مطالعه‌ای شبیه‌سازی برای ارزیابی و مقایسه مدل رگرسیون آمیخته متناهی پواسون دومتغیره در رهیافت بسامدی و بیزی صورت پذیرفته است. در بخش ۵، نحوه کاربست رهیافت بیزی پیشنهادی در قالب یک مثال کاربردی درباره عوامل موثر بر بروز سرطان‌های معده و روده بزرگ، شرح داده شده است.

۲ مدل رگرسیون آمیخته‌ی متناهی پواسون دومتغیره

در حالت تک متغیره بهترین راه حل مشکل بیش‌پراکنش در مشاهدات پاسخ جایگزین کردن توزیع دوجمله‌ای منفی به جای توزیع پواسون می‌باشد، اما در

حالت چندمتغیره استفاده از این توزیع در عمل دشوار است (برمودز و کارلیس، ۲۰۱۱). یک راه حل معقول برای رفع مشکل بیش‌پراکنش، استفاده از مدل رگرسیون پواسون با متغیر پاسخ آمیخته است. توزیع پواسون دومتغیره به‌صورت

$$BP(y_1, y_2 \mid \lambda_1, \lambda_2, \lambda_3) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{s=0}^{\min(y_1,y_2)} \binom{y_1}{s} \binom{y_2}{s} s! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^s,$$

تعریف می‌شود. بر این اساس، مدل رگرسیون پواسون دومتغیره با متغیر پاسخ آمیخته متناهی را می‌توان به‌ازای $i = 1, \cdots, n$ به‌صورت

$$\begin{aligned}
\boldsymbol{Y}_i \mid \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i} &\sim \sum_{j=1}^{g} \alpha_j BP(y_{1i}, y_{2i} \mid \lambda_{1j}(\boldsymbol{x}_{1i}), \lambda_{2j}(\boldsymbol{x}_{2i}), \lambda_{3j}(\boldsymbol{x}_{3i}))) \\
BP(\boldsymbol{y}_i|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i}) &= \sum_{j=1}^{g} [\alpha_j e^{-(\lambda_{1j}(\boldsymbol{x}_{1i})+\lambda_{2j}(\boldsymbol{x}_{2i})+\lambda_{3j}(\boldsymbol{x}_{3i}))} \\
&\times \frac{\lambda_{1j}^{y_{1i}}(\boldsymbol{x}_{1i})}{y_{1i}!} \frac{\lambda_{2j}^{y_{2i}}(\boldsymbol{x}_{2i})}{y_{2i}!} \\
&\times \sum_{s_i=0}^{\min(y_{1i}, y_{2i})} \binom{y_{1i}}{s_i} \binom{y_{2i}}{s_i} s_i! \left(\frac{\lambda_{3j}(\boldsymbol{x}_{3i})}{\lambda_{1j}(\boldsymbol{x}_{1i})\lambda_{2j}(\boldsymbol{x}_{2i})}\right)^{s_i}], \quad (1)
\end{aligned}$$

نوشت، که در آن $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})'$، $\lambda_{\ell j}(\boldsymbol{x}_{\ell i}) = \exp(\boldsymbol{x}'_{\ell i}\boldsymbol{\beta}_{\ell j})$، $\ell = 1, 2, 3$، $j = 1, \cdots, g$، $\boldsymbol{x}'_{\ell i} = (x_{1\ell i}, \cdots, x_{p\ell i})$ بردار متغیرهای تبیینی مربوط به ℓامین پارامتر توزیع پواسون دومتغیره و $\boldsymbol{\beta}'_{\ell j} = (\beta_{1\ell j}, \cdots, \beta_{p\ell j})$ بردار ضرایب رگرسیونی مربوط به ℓامین پارامتر توزیع پواسون دومتغیره و jامین زیرجامعه است. مجموعه پارامترهای مدل رگرسیون پواسون دومتغیره با متغیر پاسخ آمیخته متناهی g مؤلفه‌ای به‌صورت

$$\boldsymbol{\Phi} = \{\boldsymbol{\theta}, \boldsymbol{\alpha}\} = \{\{\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}, \boldsymbol{\beta}_{3j}\}_{j=1}^{g}, (\alpha_1, \cdots, \alpha_g)\}$$

و مشتمل بر $3gp + g$ پارامتر است. فرض کنید نمونه‌ای تصادفی به حجم n به‌صورت $\boldsymbol{\mathcal{D}} = \{(\binom{y_{1i}}{y_{2i}}, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i})\}_{i=1}^{n}$ موجود است. چون از نحوه انتساب مشاهدات به زیرجامعه‌های توزیع آمیخته اطلاعاتی در دست نیست، چنین داده‌هایی

را اصطلاحا ناقص می‌نامند و تابع درستنمایی آن‌ها به‌صورت

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\alpha}) &= \prod_{i=1}^{n}\{\sum_{j=1}^{g}\alpha_j BP(y_{1i},y_{2i}\mid \lambda_{1j}(\boldsymbol{x}_{1i}),\lambda_{2j}(\boldsymbol{x}_{2i}),\lambda_{3j}(\boldsymbol{x}_{3i}))\} \\ &= \prod_{i=1}^{n}\{\sum_{j=1}^{g}[\alpha_j e^{-(\lambda_{1j}(\boldsymbol{x}_{1i})+\lambda_{2j}(\boldsymbol{x}_{2i})+\lambda_{3j}(\boldsymbol{x}_{3i}))} \\ &\times \frac{\lambda_{1j}^{y_{1i}}(\boldsymbol{x}_{1i})}{y_{1i}!}\frac{\lambda_{2j}^{y_{2i}}(\boldsymbol{x}_{2i})}{y_{2i}!} \\ &\times \sum_{s_i=0}^{\min(y_{1i},y_{2i})}\binom{y_{1i}}{s_i}\binom{y_{2i}}{s_i}s_i!\Big(\frac{\lambda_{3j}(\boldsymbol{x}_{3i})}{\lambda_{1j}(\boldsymbol{x}_{1i})\lambda_{2j}(\boldsymbol{x}_{2i})}\Big)^{s_i}]\}, \end{aligned}$$

تعریف می‌شود. به کمک بردار متغیرهای گم‌شده $\boldsymbol{z}_i=(z_{i1},\cdots,z_{ig})^T$، که نشان‌دهنده وضعیت انتساب هر مشاهده به مؤلفه‌های توزیع آمیخته است و به‌صورت

$$z_{ij}=\begin{cases} 1 & \boldsymbol{y}_i=(y_{1i},y_{2i}) \text{ متعلق به } j \text{ امین زیرجامعه} \\ 0 & \text{در غیر این‌صورت} \end{cases}$$

تعریف می‌شود، نسخه دیگری از داده‌ها که اصطلاحا داده‌های کامل نامیده می‌شوند، ساخته می‌شود. چون هر مشاهده تنها می‌تواند به یک مؤلفه از توزیع آمیخته تعلق داشته باشد، بنا بر این روابط $\sum_{j=1}^{g} z_{ij}=1$ و $\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}=n$ برقرار است. بردارهای تصادفی $\boldsymbol{z}_i$ دارای توزیع چندجمله‌ای $\mathcal{M}(\alpha_1,\cdots,\alpha_g)$ هستند. از این رو، تابع درستنمایی داده‌های کامل به‌صورت

$$\begin{aligned} L_c(\boldsymbol{\theta},\boldsymbol{\alpha}) &= P(\boldsymbol{\mathcal{D}}\mid\boldsymbol{\theta},\boldsymbol{\alpha}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{g}\{[BP(y_{1i},y_{2i}\mid\boldsymbol{\beta}_{1j}\boldsymbol{\beta}_{2j},\boldsymbol{\beta}_{3j},\boldsymbol{x}_{1i},\boldsymbol{x}_{2i},\boldsymbol{x}_{3i})]^{z_{ij}}\alpha_j^{z_{ij}}\} \\ &= \prod_{i=1}^{n}(\prod_{j=1}^{g}[e^{-(\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j})+\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j})+\exp(\boldsymbol{x}'_{3i}\boldsymbol{\beta}_{3j}))} \\ &\times \frac{(\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j}))^{y_{1i}}}{y_{1i}!}\frac{(\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j}))^{y_{2i}}}{y_{2i}!} \\ &\times \sum_{s_i=0}^{\min(y_{1i},y_{2i})}\binom{y_{1i}}{s_i}\binom{y_{2i}}{s_i}s_i! \\ &\times \Big(\frac{\exp(\boldsymbol{x}'_{3i}\boldsymbol{\beta}_{3j})}{\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j})\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j})}\Big)^{s_i}]^{z_{ij}})\prod_{j=1}^{g}\alpha_j^{\eta_j}, \end{aligned} \qquad (۲)$$

خواهد بود، که در آن $\boldsymbol{\mathcal{D}} = \{(\binom{y_{1i}}{y_{2i}}, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i})\}_{i=1}^{n}$ نشان‌دهنده مشاهدات و $\eta_j = \sum_{i=1}^{n} z_{ij}$ بیان‌گر تعداد مشاهدات مربوط به زیرجامعه jام توزیع آمیخته است.

۱.۲ رهیافت بسامدی برازش مدل رگرسیون آمیخته پواسون دومتغیره

چون براوردگرهای ماکسیمم درستنمایی پارامترهای مدل رگرسیون پواسون دومتغیره با متغیر پاسخ آمیخته متناهی فرم بسته ندارند، یکی از روش‌های رایج برای براورد ماکسیمم درستنمایی پارامترهای این مدل، استفاده از الگوریتم EM است (دمپستر و همکاران، ۱۹۷۷). در این مدل بردار متغیرهای پنهان به‌صورت

$$\boldsymbol{b}_{ij} = (B_{1ij}, B_{2ij}, B_{3ij}) = (V_{1ij}, V_{2ij}, V_{3ij}),$$

تعریف می‌شود. با توجه به این که $\boldsymbol{y}_i$ بردار داده‌های مشاهده شده و z_i بردار متغیرهای گم‌شده است، بردار داده‌های کامل به‌صورت $\boldsymbol{y}_c = (\boldsymbol{y}_i, \mathbf{e}_{ij}, z_i)$ می‌باشد. بنابراین تابع درستنمایی داده‌های کامل به‌صورت

$$L_c(\boldsymbol{\Phi}) = \prod_{i=1}^{n}\prod_{j=1}^{g}\Big(\alpha_j BP(y_{1i}, y_{2i}|\lambda_{1j}(\boldsymbol{x}_{1i}), \lambda_{2j}(\boldsymbol{x}_{2i}), \lambda_{3j}(\boldsymbol{x}_{3i})\Big)^{z_{ij}},$$

به‌دست می‌آید، که در آن $\boldsymbol{\Phi} = (\{\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}, \boldsymbol{\beta}_{3j}\}_{j=1}^{g}, (\alpha_1, \cdots, \alpha_g))$. با فرض استقلال متغیرهای پنهان $V_{\ell ij}$، $\ell = 1, 2, 3$، $j = 1, \cdots, g$، $i = 1, \cdots, n$، لگاریتم درستنمایی داده‌های کامل برابر

$$\begin{aligned}\ell_c(\boldsymbol{\Phi}) &= \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\log\alpha_j \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\sum_{\ell=1}^{3}(-\lambda_{\ell j}(\boldsymbol{x}_{\ell i}) + v_{\ell ij}\log\lambda_{\ell j}(\boldsymbol{x}_{\ell i}) - \log v_{\ell ij}), \quad (۳)\end{aligned}$$

است. چون براورد پارامتر $\lambda_{\ell j}$ مطرح است، می‌توان رابطه (۴) را معادل

$$\ell_c(\boldsymbol{\Phi}) \propto \sum_{i=1}^{n}\sum_{j=1}^{g}(-z_{ij}\lambda_{\ell j}(\boldsymbol{x}_{\ell i}) + z_{ij}v_{\ell ij}\log\lambda_{\ell j}(\boldsymbol{x}_{\ell i})),$$

در نظر گرفت. بنابراین امید ریاضی تابع لگاریتم درستنمایی داده‌های کامل به‌صورت

$$E(\ell_c(\boldsymbol{\Phi})) = \sum_{i=1}^{n}\sum_{j=1}^{g}(-E(Z_{ij})\lambda_{\ell j}(\boldsymbol{x}_{\ell i}) + E(Z_{ij}V_{\ell ij})\log\lambda_{\ell j}(\boldsymbol{x}_{\ell i})),$$

خواهد بود. چون z_{ij} یک متغیر تصادفی برنولی است، وقتی iامین مشاهده به jامین مؤلفه تعلق نداشته باشد، $v_{\ell ij}$ دارای مقدار صفر و در غیر این صورت دارای مقداری غیر صفر می‌باشد. بنا بر این می‌توان نوشت

$$E(Z_{ij}V_{\ell ij}) = p(z_{ij})E(V_{\ell ij} \mid z_{ij} = 1). \qquad (۴)$$

امید ریاضی سمت راست رابطه (۵)، امید ریاضی متغیرهای پنهان به شرط تعلق iامین مشاهده به jامین مؤلفه توزیع آمیخته است. بنا بر این در مرحله E، تنها محاسبه $E(Z_{ij}|\boldsymbol{y}_i, \lambda_{1j}(\boldsymbol{x}_{1i}), \lambda_{2j}(\boldsymbol{x}_{2i}), \lambda_{3j}(\boldsymbol{x}_{3i}))$ و $E(V_{\ell ij}|\boldsymbol{y}_i, z_{ij} = 1, \lambda_{1j}(\boldsymbol{x}_{1i}), \lambda_{2j}(\boldsymbol{x}_{2i}), \lambda_{3j}(\boldsymbol{x}_{3i}))$ لازم است. به طور کلی مراحل الگوریتم EM به‌صورت زیر است:

۱) **مقدار اولیه**. مقادیر اولیه مناسبی را برای پارامترهای مدل در نظر بگیرید.

۲) **مرحله** E. در مرحله tام مقادیر

$$\begin{aligned}
w_{ij}^{(t)} &= E(Z_{ij} \mid \boldsymbol{y}_i, \lambda_j^{(t)}(\boldsymbol{x}_{1i}), \lambda_{2j}^{(t)}(\boldsymbol{x}_{2i}), \lambda_{3j}^{(t)}(\boldsymbol{x}_{3i})) \\
&= \frac{\alpha_j^{(t)} P(\boldsymbol{y}_i \mid \lambda_{1j}^{(t)}(\boldsymbol{x}_{1i}), \lambda_{2j}^{(t)}(\boldsymbol{x}_{2i}), \lambda_{3j}^{(t)}(\boldsymbol{x}_{3i}))}{\sum_{j=1}^{g}\alpha_j^{(t)} P(\boldsymbol{y}_i \mid \lambda_{1j}^{(t)}(\boldsymbol{x}_{1i}), \lambda_{2j}^{(t)}(\boldsymbol{x}_{2i}), \lambda_{3j}^{(t)}(\boldsymbol{x}_{3i}))}, \\
b_{1ij}^{(t)} &= E(V_{1ij} \mid \boldsymbol{y}_i, z_{ij} = 1, \lambda_{1j}^{(t)}(\boldsymbol{x}_{1i}), \lambda_{2j}^{(t)}(\boldsymbol{x}_{2i}), \lambda_{3j}^{(t)}(\boldsymbol{x}_{3i})) \\
&= y_{1i} - b_{3ij}^{(t)}, \\
b_{2ij}^{(t)} &= E(V_{2ij} \mid \boldsymbol{y}_i, z_{ij} = 1, \lambda_{1j}^{(t)}(\boldsymbol{x}_{1i}), \lambda_{2j}^{(t)}(\boldsymbol{x}_{2i}), \lambda_{3j}^{(h)}(\boldsymbol{x}_{3i})) \\
&= y_{2i} - b_{3ij}^{(t)},
\end{aligned}$$

را محاسبه کنید، که در آن‌ها

$$b_{3ij}^{(t)} = E(V_{3ij} \mid \boldsymbol{y}_i, z_{ij} = 1, \lambda_{1j}^{(t)}(\boldsymbol{x}_{1i}), \lambda_{2j}^{(t)}(\boldsymbol{x}_{2i}), \lambda_{3j}^{(t)}(\boldsymbol{x}_{3i}))$$

$$= \begin{cases} \frac{BP[y_{1i}-1, y_{2i}-1 | \lambda_{1j}^{(t)}(\boldsymbol{x}_{1i}), \lambda_{2j}^{(t)}(\boldsymbol{x}_{2i}), \lambda_{3j}^{(t)}(\boldsymbol{x}_{3i})]}{BP[y_{1i}, y_{2i} | \lambda_{1j}^{(t)}(\boldsymbol{x}_{1i}), \lambda_{2j}^{(t)}(\boldsymbol{x}_{2i}), \lambda_{3j}^{(t)}(\boldsymbol{x}_{3i})]} \lambda_{3j}^{(t)}(\boldsymbol{x}_{3i}) & y_{1i}y_{2i} > 0 \\ 0 & y_{1i}y_{2i} = 0. \end{cases}$$

۳) **مرحله** M. براساس مقادیر پارامترها در مرحله tام، براورد پارامترها در مرحله $(t+1)$ام را به‌صورت

$$\begin{aligned} \alpha_j^{(t+1)} &= \frac{\sum_{i=1}^n w_{ij}^{(t)}}{n} \\ \boldsymbol{\beta}_{1j}^{(t+1)} &= \hat{\boldsymbol{\beta}}(\boldsymbol{y}_1 - \boldsymbol{b}_{3j}^{(t)}, \boldsymbol{x}_1, \boldsymbol{w}_j^{(t)}), \\ \boldsymbol{\beta}_{2j}^{(t+1)} &= \hat{\boldsymbol{\beta}}(\boldsymbol{y}_2 - \boldsymbol{b}_{3j}^{(t)}, \boldsymbol{x}_2, \boldsymbol{w}_j^{(t)}), \\ \boldsymbol{\beta}_{3j}^{(t+1)} &= \hat{\boldsymbol{\beta}}(\boldsymbol{b}_{3j}^{(t)}, \boldsymbol{x}_3, \boldsymbol{w}_j^{(t)}). \end{aligned}$$

محاسبه کنید، که در آن $\boldsymbol{b}_{3j} = [b_{1j}, \cdots, b_{nj}]^T$ یک بردار $n \times 1$ بعدی و $\hat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w})$ براوردگر ماکسیمم درستنمایی موزون پارامترهای یک مدل پواسون دومتغیره با بردار پاسخ $\boldsymbol{y}$ هستند.

۴) **همگرایی**. مراحل E و M را تا همگرا شدن الگوریتم تکرار کنید.

انتخاب مقادیر اولیه مناسب و تعیین قاعده همگرایی الگوریتم، دو مساله مهم و تاثیرگذار بر دقت براوردهای حاصل از الگوریتم EM هستند. برای بررسی همگرایی الگوریتم EM قواعد مختلفی پیشنهاد شده است. به‌عنوان نمونه می‌توان از قاعده ساده‌ای به‌صورت

$$|\boldsymbol{\Phi}^{(t+1)} - \boldsymbol{\Phi}^{(t)}| \leq \xi,$$

استفاده کرد، که در آن ξ کمیت کوچکی است و آستانه همگرایی نامیده می‌شود (بوهنینگ و همکاران، ۱۹۹۴). انتخاب مقدار اولیه بر نرخ همگرایی الگوریتم تاثیرگذار است و اگر مقادیر اولیه نامناسبی اختیار شوند، الگوریتم همگرایی کندی خواهد داشت. معمولا توصیه می‌شود مقادیر اولیه مختلفی مورد آزمایش قرار گیرند تا بهترین نرخ همگرایی و برازش حاصل شود (دمپستر و همکاران، ۱۹۷۷).

به‌طور کلی، هنگامی که مدل پیچیده و تعداد پارامترها زیاد باشد، تعیین مقادیر اولیه مناسب برای پارامترها کار ساده‌ای نیست. از طرفی، چون خواص بهینگی براوردگرهای ماکسیمم درستنمایی تنها برای نمونه‌های بزرگ بروز می‌کند، مدل بسامدی برای نمونه‌های کوچک از کارایی مطلوبی برخوردار نیست. از این رو، در بخش بعد توسعه رهیافت بیزی مساله به‌عنوان مدل جایگزین رهیافت بسامدی برمودز و کارلیس (۲۰۱۲) مطرح شده است.

۳ رهیافت بیزی برازش مدل رگرسیون آمیخته پواسون دو متغیره

به‌طور کلی تحلیل بیزی مدل‌های آمیخته دارای پیچیدگی‌های زیاد و دشواری‌های خاص خود است. توزیع پسین در این مدل‌ها معمولا پیچیده و فاقد فرم بسته است. به‌علاوه نمونه‌گیری مناسب از توزیع پسین اغلب امری دشوار و زمان‌بر است (مارین و همکاران، ۲۰۰۵). براورد بیزی پارامترها در مدل‌های آمیخته بر اساس تابع درستنمایی داده‌های کامل صورت می‌پذیرد و از این نظر شبیه رهیافت بسامدی و استفاده از الگوریتم EM است.

۱.۳ توزیع‌های پیشین

برای مدل رگرسیون پواسون دومتغیره با متغیر پاسخ آمیخته متناهی g مؤلفه‌ای، توزیع پیشین پارامترها را می‌توان با فرض استقلال پیشینی ضرایب رگرسیونی و ضرایب آمیختگی به‌صورت

$$\pi(\boldsymbol{\Phi}) = \prod_{j=1}^{g} [\pi(\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}, \boldsymbol{\beta}_{3j})] \pi(\alpha_1, \cdots, \alpha_g), \qquad (۵)$$

نوشت. معمولا برای ضرایب آمیختگی، $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_g)$، توزیع پیشین دیریکله به‌صورت

$$\begin{aligned} \pi(\alpha_1, \cdots, \alpha_g) &= Dirichlet(c_1, \cdots, c_g) \\ &= \frac{1}{B(c)} \prod_{j=1}^{g} \alpha_j^{c_j - 1} \end{aligned} \qquad (۶)$$

در نظر گرفته می‌شود. این انتخاب در ادبیات مدل‌های آمیخته، یکی از انتخاب‌های مرسوم و کلاسیک است (فرورس-شناتر، ۲۰۰۶). از آن‌جا که توزیع دیریکله تعمیم‌یافته توزیع بتا است، بر اساس رابطه‌ای که بین توزیع بتا و $u(۰, ۱)$ وجود دارد و به‌منظور پرهیز از پیش داوری معمولا ابرپارامترهای توزیع پیشین به‌صورت $c_۰ = c_۱ = c_۲ = \cdots = c_g = c_۰$ و برابر در نظر گرفته می‌شود. اگر $c_۰ = ۱$ باشد، آن‌گاه توزیع یکنواخت d-بعدی، $u(۰, ۱)^d$، حاصل می‌شود. البته فرض برابری ابرپارامترهای توزیع دیریکله زمانی مطلوب است که همه ضرایب آمیختگی با هم برابر باشند، به این معنی که تحلیل‌گر معتقد باشد، همه مؤلفه‌های توزیع آمیخته به یک اندازه در نمونه مشاهده شده سهم دارند. از این رو، باید تعلق مشاهدات به هریک از زیرجوامع را در نظر گرفت. یک روش برای این منظور در حالت‌های یک و دومتغیره رسم نمودار مشاهدات و مقایسه فراوانی مشاهدات در مؤلفه‌های تشکیل‌دهنده توزیع آمیخته است. البته این کار در حالت‌های چندمتغیره دشوار است. به‌عنوان مثال، اگر در یک مدل آمیخته دو مؤلفه‌ای، فراوانی مشاهدات مربوط به مؤلفه اول از فراوانی مشاهدات مربوط به مؤلفه دوم کم‌تر (بیش‌تر) باشد، ابرپارامترهای توزیع دیریکله را باید به نحوی تعیین کرد که این توزیع چوله به راست (چپ) باشد.

بسیاری از محققان، از جمله گلمن و همکاران (۲۰۰۸)، استفاده از توزیع پیشین نرمال برای ضرایب مدل‌های رگرسیونی را در حالت کلی منطقی می‌دانند. بر این اساس، برای ضرایب رگرسیونی $(\beta_{۱j}, \beta_{۲j}, \beta_{۳j})$، $j = ۱, \cdots, g$، توزیع نرمال چندمتغیره به‌صورت

$$\pi(\boldsymbol{\beta}_{۱j}, \boldsymbol{\beta}_{۲j}, \boldsymbol{\beta}_{۳j}) \sim N_{۳p}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad j = ۱, \cdots, g. \qquad (۷)$$

به‌عنوان توزیع پیشین در نظر گرفته شده است. بردار میانگین و ماتریس کواریانس توزیع پیشین باید طوری در نظر گرفته شوند که حاوی اطلاعات مناسبی در مورد پارامترها باشند و توزیع پسین را به سمت برازش مناسب مدل هدایت کنند. با فرض مستقل بودن ضرایب رگرسیونی، ماتریس کواریانس توزیع پیشین قطری است. اکنون با انتخاب توزیع پیشین مناسب برای پارامترهای مدل و محاسبه توزیع

پسین، می‌توان به استنباط پسینی پیرامون پارامترها پرداخت.

۲.۳ توزیع پسین

با توجه به توزیع‌های پیشین (۷) و (۷)، توزیع پسین به‌صورت

$$\begin{aligned}
\pi(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\alpha}\mid\boldsymbol{\mathcal{D}}) &\propto p(\boldsymbol{\mathcal{D}}\mid\boldsymbol{\theta},\boldsymbol{\alpha})\cdot\pi(\boldsymbol{\theta})\cdot\pi(\boldsymbol{\alpha})\\
&= \prod_{i=1}^{n}\{\prod_{j=1}^{g}[e^{-(\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j})+\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j})+\exp(\boldsymbol{x}'_{3i}\boldsymbol{\beta}_{3j}))}\\
&\times \frac{(\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j}))^{y_{1i}}}{y_{1i}!}\frac{(\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j}))^{y_{2i}}}{y_{2i}!}\\
&\times \sum_{s_i=0}^{\min(y_{1i},y_{2i})}\binom{y_{1i}}{s_i}\binom{y_{2i}}{s_i}s_i!\\
&\times \Big(\frac{\exp(\boldsymbol{x}'_{3i}\boldsymbol{\beta}_{3j})}{\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j})\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j})}\Big)^{s_i}\Big]^{z_{ij}}\alpha_j^{z_{ij}}\}\\
&\times \prod_{j=1}^{g}[\pi(\boldsymbol{\beta}_{1j},\boldsymbol{\beta}_{2j},\boldsymbol{\beta}_{3j})](\prod_{j=1}^{g}\alpha_j^{c_j-1}),
\end{aligned}\tag{۸}$$

به‌دست می‌آید. به‌دلیل پیچیدگی توزیع پسین، امکان محاسبه ثابت چگالی‌ساز عملا وجود ندارد. از این رو، توزیع پسین فاقد شکل بسته و شناخته‌شده است. بر این اساس، می‌توان از زنجیر مارکف مونت کارلویی برای نمونه‌گیری از این توزیع پسین استفاده کرد. سپس نمونه‌های حاصل را برای هر گونه استنباط پسینی پیرامون پارامترها به کار گرفت.

۱.۲.۳ توزیع‌های پسین شرطی کامل

در این بخش نحوه کاربست الگوریتم گیبز برای نمونه‌گیری از توزیع پسین (۸) شرح داده می‌شود. لازمه استفاده از الگوریتم گیبز شناخت توزیع‌های پسین شرطی کامل است. بنابراین در این بخش توزیع‌های پسین شرطی کامل برای ضرایب آمیختگی، ضرایب رگرسیون و متغیرهای گم‌شده z_i محاسبه می‌شوند. بر اساس

توزیع پسین (۸) و توزیع پیشین ضرایب آمیختگی در (۷)، توزیع پسین شرطی کامل ضرایب آمیختگی به‌صورت

$$\begin{aligned}\pi(\boldsymbol{\alpha} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\mathcal{D}}) &\propto \prod_{j=1}^{g} \alpha_j^{c_j - 1} \prod_{j=1}^{g} \alpha_j^{\sum_{i=1}^{n} z_{ij}} \\ &= \prod_{j=1}^{g} \alpha_j^{c_j + \eta_j - 1},\end{aligned}$$

محاسبه می‌شود. بنابراین، توزیع پسین شرطی کامل برای ضریب آمیختگی، توزیع دیریکله به‌صورت

$$\pi(\boldsymbol{\alpha} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\mathcal{D}}) = Dirichlet(c_1 + \eta_1, \cdots, c_g + \eta_g),$$

است. به‌همین ترتیب توزیع پسین شرطی کامل ضرایب رگرسیونی $\boldsymbol{\theta} = \{\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}, \boldsymbol{\beta}_{3j}\}_{j=1}^{g}$ به‌صورت

$$\begin{aligned}\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\mathcal{D}}) &= \int_0^1 \cdots \int_0^1 \pi(\boldsymbol{\theta}, \boldsymbol{\alpha} \mid \boldsymbol{Z}, \boldsymbol{\mathcal{D}}) d\alpha_1 \cdots d\alpha_g \\ &\propto \prod_{i=1}^{n} \Big(\prod_{j=1}^{g} \Big[e^{-(\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j}) + \exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j}) + \exp(\boldsymbol{x}'_{3i}\boldsymbol{\beta}_{3j}))} \\ &\quad \times \frac{(\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j}))^{y_{1i}}}{y_{1i}!} \frac{(\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j}))^{y_{2i}}}{y_{2i}!} \\ &\quad \times \sum_{s_i=0}^{\min(y_{1i}, y_{2i})} \binom{y_{1i}}{s_i} \binom{y_{2i}}{s_i} s_i! \\ &\quad \times \Big(\frac{\exp(\boldsymbol{x}'_{3i}\boldsymbol{\beta}_{3j})}{\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j}) \exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j})} \Big)^{s_i} \Big]^{z_{ij}} \Big) \\ &\quad \times \prod_{j=1}^{g} \varphi_{3p}((\boldsymbol{\beta}_{1j},\ \boldsymbol{\beta}_{2j},\ \boldsymbol{\beta}_{3j}); \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),\end{aligned}$$

است، که در آن $\varphi_{3p}(\cdot,\cdot,\cdot)$ تابع چگالی توزیع نرمال $3p-$ بعدی است. توزیع پسین شرطی کامل متغیر نشان‌گر z_{ij} که وضعیت تعلق iامین مشاهده به jامین مؤلفه را مشخص می‌سازد، به‌صورت

$$P(z_{ij} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}) = \frac{\alpha_j BP(y_{1i}, y_{2i} \mid \boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}, \boldsymbol{\beta}_{3j}, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i})}{\sum_{j=1}^{g} \alpha_j BP(y_{1i}, y_{2i} \mid \boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}, \boldsymbol{\beta}_{3j}, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i})} \quad (۹)$$

است. بر اساس احتمال‌های پسین (۹)، توزیع پسین شرطی کامل $z_i = (z_{i1}, \cdots, z_{ig})$، توزیع g - جمله‌ای به‌صورت

$$\pi(z_i \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}) = \mathcal{M}(1, Pr(z_{i1} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}), \cdots, Pr(z_{ig} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}})),$$

است. بنابراین، توزیع پسین شرطی کامل متغیرهای گم‌شده، ضرایب آمیختگی و ضرایب رگرسیونی، به‌ترتیب عبارتند از

$$\begin{aligned}
\pi(z_i \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}) &= \mathcal{M}(1, Pr(z_{i1} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}), \cdots, Pr(z_{ig} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}})),\\
\pi(\boldsymbol{\alpha} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\mathcal{D}}) &= Dirichlet(c_1 + \eta_1, \cdots, c_g + \eta_g),\\
\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\mathcal{D}}) &\propto \prod_{i=1}^{n}\Big(\prod_{j=1}^{g}\big[e^{-(\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j})+\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j})+\exp(\boldsymbol{x}'_{3i}\boldsymbol{\beta}_{3j}))}\\
&\quad\times \frac{(\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j}))^{y_{1i}}}{y_{1i}!}\frac{(\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j}))^{y_{2i}}}{y_{2i}!}\\
&\quad\times \sum_{s_i=0}^{\min(y_{1i},y_{2i})}\binom{y_{1i}}{s_i}\binom{y_{2i}}{s_i}s_i!\\
&\quad\times \Big(\frac{\exp(\boldsymbol{x}'_{3i}\boldsymbol{\beta}_{3j})}{\exp(\boldsymbol{x}'_{1i}\boldsymbol{\beta}_{1j})\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_{2j})}\Big)^{s_i}\big]^{z_{ij}}\Big)\\
&\quad\times \prod_{j=1}^{g}\varphi_{3p}\Big((\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j}, \boldsymbol{\beta}_{3j}); \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\Big).
\end{aligned}$$

اکنون با در دست داشتن توزیع‌های شرطی کامل می‌توان از توزیع پسین توأم $\{\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{\mathcal{D}}\}$ به کمک الگوریتم نمونه برداری گیبز[۳] نمونه‌گیری نمود. همان‌طور که مشاهده می‌شود توزیع پسین شرطی کامل $\{\boldsymbol{\theta} \mid \boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}\}$ صورت بسته و شناخته شده‌ای ندارد، بنابراین با استفاده از الگوریتم متروپولیس - هستینگس[۴] از توزیع پسین شرطی کامل $\{\boldsymbol{\theta} \mid \boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}\}$ نمونه تولید می‌شود. برای نمونه‌گیری از توزیع پسین به تلفیقی از الگوریتم‌های گیبز و متروپلیس - هاستینگز نیاز است. مراحل کلی الگوریتم گیبز برای نمونه‌گیری از توزیع پسین (۸) به شرح زیر است:

۱) مقدار اولیه‌ی $\boldsymbol{\alpha}^{(0)}$ را برای $\boldsymbol{\alpha}$ و $\boldsymbol{\theta}^{(0)} = \left\{\boldsymbol{\beta}_{1j}^{(0)}, \boldsymbol{\beta}_{2j}^{(0)}, \boldsymbol{\beta}_{3j}^{(0)}\right\}_{j=1}^{g}$ را برای $\boldsymbol{\theta}$ در نظر بگیرید.

---

[۳] Gibbs sampling

[۴] Metropolis-Hastings

۲) در مرحله $(t+1)$ام بر اساس مقادیر حاصل از مرحله‌ی tام، $\boldsymbol{Z}$، $\boldsymbol{\alpha}$ و $\boldsymbol{\theta}$ را به‌صورت زیر به‌روز کنید:

$$\pi(Z_i^{(t+1)} \mid \theta^{(t)}, \alpha^{(t)}, \mathcal{D}) = \mathcal{M}(1, P(z_{i1}^{(t+1)} \mid \theta^{(t)}, \alpha^{(t)}, \mathcal{D}), \cdots, P(z_{ig}^{(t+1)} \mid \theta^{(t)}, \alpha^{(t)}, \mathcal{D})),$$

$$\pi(\alpha^{(t+1)} \mid z^{(t+1)}, \theta^{(t)}, \mathcal{D}) = Dirichlet(c_1 + \eta_1^{(t+1)}, \cdots, c_g + \eta_g^{(t+1)}),$$

$$\pi(\theta^{(t+1)} \mid \alpha^{(t+1)}, z^{(t+1)}, \mathcal{D}) \propto \prod_{i=1}^{n}(\prod_{j=1}^{g}[e^{-(\exp(x'_{1i}\beta_{1j}^{(t)})+\exp(x'_{2i}\beta_{2j}^{(t)})+\exp(x'_{3i}\beta_{3j}^{(t)}))}$$
$$\times \quad \frac{(\exp(x'_{1i}\beta_{1j}^{(t)}))^{y_{1i}}}{y_{1i}!}\frac{(\exp(x'_{2i}\beta_{2j}^{(t)}))^{y_{2i}}}{y_{2i}!}\sum_{s_i=0}^{\min(y_{1i},y_{2i})}\binom{y_{1i}}{s_i}\binom{y_{2i}}{s_i}s_i!$$
$$\times \quad \frac{\exp(x'_{3i}\beta_{3j}^{(t)})}{\exp(x'_{1i}\beta_{1j}^{(t)})\exp(x'_{2i}\beta_{2j}^{(t)})})^{s_i}]^{z_{ij}^{(t+1)}}\prod_{j=1}^{g}\varphi_{3p}((\beta_{1j}^{(t)}, \beta_{2j}^{(t)}, \beta_{3j}^{(t)}); \mu_j, \Sigma_j).$$

۳) مرحله ۲ را تا رسیدن الگوریتم به حالت مانایی تکرار کنید.

در مرحله ۲ الگوریتم، برای نمونه‌گیری از توزیع پسین شرطی کامل $\{\boldsymbol{\theta} \mid \boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}\}$ به‌دلیل بسته نبودن شکل توزیع، از الگوریتم متروپولیس - هستینگس برای نمونه‌گیری از این توزیع استفاده می‌شود.

۴ مطالعه شبیه‌سازی

در این بخش مدل بیزی پیشنهادی که در بخش ۳ مورد بحث قرار گرفت، در چارچوب یک مطالعه‌ی شبیه‌سازی مورد ارزیابی قرار گرفته و کارایی آن با مدل بسامدی معرفی شده توسط برمودز و کارلیس (۲۰۱۲) مقایسه می‌شود. برای این منظور از یک طرح شبیه‌سازی مشابه آن‌چه در جرج و مکلاک (۱۹۹۳) و رفتری و همکاران (۱۹۹۷) برای ارزیابی مدل‌های رگرسیونی توصیه شده است، استفاده می‌شود. فرض کنید $g = p = 2$. از این رو فضای پارامتری شامل ۱۳ پارامتر است و در این مطالعه مقادیر ضرایب رگرسیونی به‌صورت

$$\begin{aligned}\boldsymbol{\theta} &= \{\beta_{110}, \beta_{111}, \beta_{210}, \beta_{211}, \beta_{310}, \beta_{311}, \beta_{120}, \beta_{121}, \beta_{220}, \beta_{221}, \beta_{320}, \beta_{321}\} \\ &= \{0/4, 0/4, 0/4, 0/4, 0/5, 0/5, 0/8, 0/8, 0/8, 0/8, 0/9, 0/9\}\end{aligned}$$

در نظر گرفته شده‌اند. به‌علاوه، به‌منظور ارزیابی نحوه اثرگذاری فراوانی مشاهدات مربوط به مؤلفه‌های تشکیل‌دهنده توزیع آمیخته بر کارایی مدل‌های مورد بحث، برای ضریب آمیختگی α مقادیر مختلف (۰/۱, ۰/۲۵, ۰/۵, ۰/۷۵, ۰/۹۹) در نظر گرفته شده است. مقادیر متغیرهای تبیینی x_{1i}، x_{2i} و x_{3i}، $i = 1, \cdots, n$، از توزیع نرمال $N(\mu, \sigma^2)$ شبیه‌سازی شده‌اند. بر این اساس، با توجه به نرمال بودن ترکیب خطی متغیرهای نرمال و رابطه $\lambda(\boldsymbol{x}) = \exp(\boldsymbol{x}'\boldsymbol{\beta})$، متغیر تصادفی $\lambda(\boldsymbol{X})$ دارای توزیع لگ نرمال $LN(\mu, \sigma^2)$ است. سپس مقادیر پاسخ $\boldsymbol{y}_i = \binom{y_{1i}}{y_{2i}}$، از طریق رابطه (۲) شبیه‌سازی شده است. در نهایت با فرض موجود بودن مشاهدات $\{\binom{y_{1i}}{y_{2i}}, \boldsymbol{x}_i\}_{i=1}^n$ ضرایب رگرسیونی با رهیافت بسامدی (برمودز و کارلیس، ۲۰۱۲) و هم‌چنین رهیافت بیزی پیشنهاد شده در این مقاله، براورد و میزان کارایی نسبی دو رهیافت در براورد ضرایب رگرسیونی و پارامتر آمیختگی بر اساس معیار کمترین توان‌های دوم خطا[۵] (MSE) به‌صورت

$$E(\hat{\boldsymbol{\beta}}_{EM}, \hat{\boldsymbol{\beta}}_B) = \frac{MSE(\hat{\boldsymbol{\beta}}_B)}{MSE(\hat{\boldsymbol{\beta}}_{ML})},$$
$$E(\hat{\alpha}_{EM}, \hat{\alpha}_B) = \frac{MSE(\hat{\alpha}_B)}{MSE(\hat{\alpha}_{ML})},$$

به‌دست آمده‌اند. در تمامی محاسبات به‌منظور لحاظ نمودن تغییرپذیری مشاهدات شبیه‌سازی شده، تعداد تکرار برابر با ۵۰۰ در نظر گرفته شده است. نتایج به‌ازای مقادیر مختلف α و اندازه‌های نمونه‌ای متفاوت در جداول ۱ و ۲ ارائه شده‌اند. گرچه همگرایی یک زنجیر مارکف به توزیع مانای خود تحت شرایط معین تضمین شده است، اما چون هیچ راه نظری برای تعیین زمان همگرایی یک زنجیر مارکف وجود ندارد، معمولا از نمودارهای اثر و خودهمبستگی برای تشخیص همگرایی استفاده می‌شود. به‌علاوه، برخی معیارهای عددی برای اطمینان بیش‌تر از همگرایی زنجیرهای مارکف، توسط محققانی مانند گلمن و روبین (۱۹۹۲) پیشنهاد شده و بسته‌های نرم‌افزاری مناسبی برای این منظور تهیه شده است. در این مقاله از تابع $gelman.diag$ در بسته $Coda$ از نرم‌افزار R برای تشخیص همگرایی زنجیرهای مارکف استفاده شده است. نمودارهای اثر و خودهمبستگی برای یک نمونه نوعی

[۵] Mean square errors

شبیه‌سازی شده از توزیع پسین ضرایب رگرسیونی، در شکل ۱ رسم شده‌اند. این نمودارها به خوبی همگرایی زنجیرهای مارکف تولید شده از توزیع‌های پسین ضرایب رگرسیونی را نشان می‌دهند. البته با توجه به پیچیدگی مدل، همگرایی زنجیرهای مارکف خصوصا برای نمونه‌های بزرگ، زمان‌بر است.

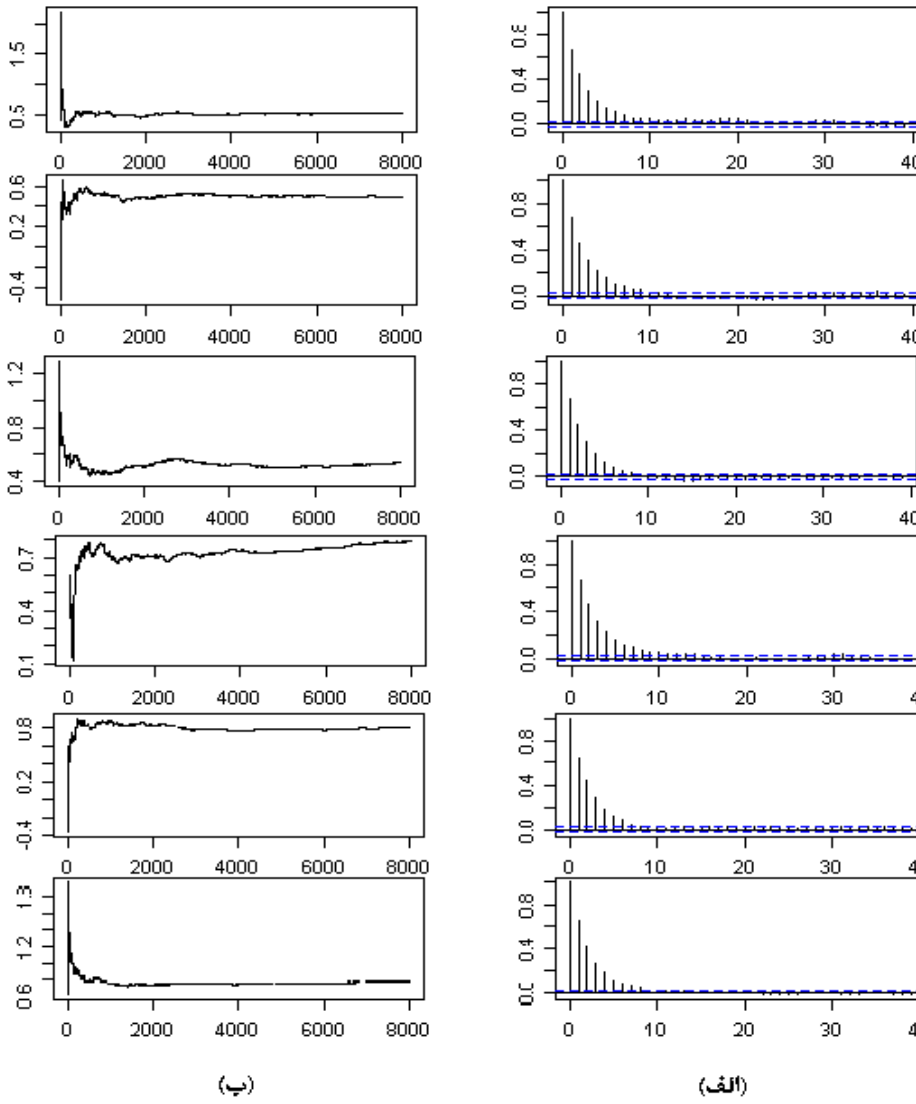
همان‌طور که ملاحظه می‌شود، میزان کارایی رهیافت بیزی در براورد ضرایب رگرسیونی به‌ازای تمام مقادیر α و تمام اندازه نمونه‌های در نظر گرفته شده، از کارایی رهیافت بسامدی بیش‌تر است. برتری کارایی براوردگر حاصل از مدل بیزی به براوردگر بسامدی ماکسیمم درستنمایی به‌ویژه برای مقادیر ضریب آمیختگی نزدیک به ۰/۵ کاملا قابل توجه است. البته چنان‌چه انتظار می‌رود، کارایی براوردگرهای ماکسیمم درستنمایی بر اساس خواص مجانبی این نوع براوردگرها با افزایش اندازه نمونه بهبود می‌یابد. در مورد پارامتر ضریب آمیختگی، α، که در تحلیل رگرسیونی مستقیما مورد علاقه نیست، وضعیت تا حدودی متفاوت است. به‌گونه‌ای که رهیافت بیزی تنها به‌ازای مقادیر متوسط α برای تمام اندازه‌های نمونه‌ای از رهیافت بسامدی کاراتر است (ردیف سوم جدول ۲). برای مقادیر αی نزدیک به ۰ یا ۱ برتری کارایی رهیافت بیزی به بسامدی تنها برای نمونه‌های کوچک $\{۵۰, ۲۵\}$ صادق است و براوردگر ماکسیمم درستنمایی ضریب آمیختگی برای نمونه‌های بزرگ بر اساس خواص مجانبی خود، کارایی بیش‌تری دارد. موارد نادری ناپایداری در نتایج نیز مشاهده می‌شود که ناشی از حجم نمونه‌های کوچک و مشاهدات شبیه‌سازی شده است. تاکید بر نمونه‌های کوچک در مطالعه شبیه‌سازی به این دلیل است که بر اساس اصول استنباط بیزی، برای اندازه نمونه‌های بزرگ درستنمایی بر پیشین غلبه می‌کند و عملا تفاوتی بین نتایج حاصل از تحلیل این مدل از دو دیدگاه بسامدی و بیزی وجود نخواهد داشت. از این رو، برای نمونه‌های بزرگ می‌توان مدل بسامدی و استفاده از الگوریتم EM را از نقطه نظر سادگی و سرعت عمل بیش‌تر، توصیه نمود.

جدول ۱: کارایی نسبی براوردگر بسامدی نسبت به براوردگر بیزی برای ضرایب رگرسیون به‌ازای مقادیر مختلف ضرایب آمیختگی مولفه اول

$\beta_{\ell jk}$	α	حجم نمونه			
		۲۵	۵۰	۱۰۰	۲۰۰
$\beta_{۱۱۰}$	۰/۱	۰/۰۳۳۴	۰/۰۴۹۸	۰/۱۳۷۹	۰/۰۸۷۷
	۰/۲۵	۰/۰۲۶۶	۰/۰۵۱۷	۰/۱۱۷۷	۰/۴۳۴۳
	۰/۵	۰/۰۱۹۳	۰/۰۵۱۷	۰/۰۷۹۸	۰/۰۱۹۳
	۰/۷۵	۰/۱۶۹۹	۰/۲۱۶۵	۰/۱۱۸۰	۰/۱۶۹۹
	۰/۹۹	۰/۰۷۹۷	۰/۰۶۵۶	۰/۱۶۱۷	۰/۰۷۹۷
$\beta_{۱۱۱}$	۰/۱	۰/۰۲۲۵	۰/۶۴۹۵	۰/۰۵۶۹	۰/۳۷۸۸
	۰/۲۵	۰/۰۲۰۱	۰/۵۸۳۲	۱/۷۱۸۴	۰/۰۸۷۳
	۰/۵	۰/۰۵۹۴	۰/۶۱۹۷	۰/۱۷۰۹	۰/۶۵۸۶
	۰/۷۵	۰/۳۳۰۸	۰/۰۹۲۷	۰/۱۱۵۶	۰/۲۲۷۶
	۰/۹۹	۰/۱۷۹۱	۱/۱۳۲۳	۰/۱۰۳۲	۰/۲۱۸۶
$\beta_{۲۱۰}$	۰/۱	۰/۱۲۳۲	۰/۰۴۰۸	۰/۲۴۰۱	۰/۱۰۵۰
	۰/۲۵	۰/۰۷۹۹	۰/۰۴۲۶۷	۰/۲۴۵۸	۰/۲۱۸۳
	۰/۵	۰/۰۵۴۴	۰/۰۴۱۳	۰/۰۶۸۴	۰/۱۳۷۱
	۰/۷۵	۰/۱۲۷۶	۰/۰۸۷۹	۰/۲۲۶۹	۰/۹۴۱۲
	۰/۹۹	۰/۰۷۶۸	۰/۰۷۵۶	۰/۶۹۰۹	۰/۴۳۳۱
$\beta_{۲۱۱}$	۰/۱	۰/۰۵۶۲	۰/۱۱۶۶	۰/۰۵۰۷	۰/۰۷۱۰
	۰/۲۵	۰/۰۷۳۷	۰/۱۲۶۶	۰/۱۰۲۵	۰/۲۹۹۹
	۰/۵	۰/۰۵۳۳	۰/۱۱۸۲	۰/۰۸۲۳	۰/۴۲۸۱
	۰/۷۵	۰/۱۱۹۸	۰/۲۵۲۱	۰/۰۸۹۴	۰/۹۴۷۷
	۰/۹۹	۰/۲۰۲۹	۰/۲۳۰۳	۰/۱۲۶۴	۰/۱۴۳۶
$\beta_{۳۱۰}$	۰/۱	۰/۰۲۹۷	۰/۰۴۳۹	۱/۴۲۲۵	۰/۱۳۵۵
	۰/۲۵	۰/۰۳۱۲	۰/۰۴۴۶	۰/۸۴۲۴	۰/۰۸۳۵
	۰/۵	۰/۰۲۴۴	۰/۰۴۳۸	۰/۰۲۳۹	۰/۱۸۶۱
	۰/۷۵	۰/۰۲۷۴	۰/۰۶۲۰	۰/۱۰۳۳	۰/۱۳۵۲
	۰/۹۹	۰/۰۱۷۲	۰/۰۶۴۰	۰/۲۴۳۴	۰/۱۶۶۴
$\beta_{۳۱۱}$	۰/۱	۰/۶۳۸۶	۰/۳۸۱۱	۰/۱۷۸۳	۰/۲۰۱۱
	۰/۲۵	۰/۹۰۳۵	۰/۳۵۷۷	۰/۱۲۲۹	۰/۱۰۵۴
	۰/۵	۰/۰۴۶۲	۰/۳۷۳۲	۰/۰۷۸۰	۰/۱۳۷۴
	۰/۷۵	۰/۰۴۷۲	۰/۱۷۶۷	۰/۱۵۶۹	۰/۱۰۰۴
	۰/۹۹	۰/۰۵۲۱	۰/۳۱۵۳	۰/۸۵۹۹	۰/۱۳۷۹

جدول ۲: کارایی نسبی براوردگر بسامدی نسبت به براوردگر بیزی برای ضرایب رگرسیون به‌ازای مقادیر مختلف ضرایب آمیختگی مولفه دوم

$\beta_{\ell jk}$	α	حجم نمونه			
		۲۵	۵۰	۱۰۰	۲۰۰
$\beta_{۱۲۰}$	۰/۱	۱/۲۰۶۷	۰/۱۲۹۰	۰/۱۷۵۱	۰/۱۵۳۱
	۰/۲۵	۰/۲۳۴۰	۰/۱۲۲۰	۰/۱۵۱۶	۰/۰۶۹۳
	۰/۵	۰/۱۵۲۹	۰/۱۱۹۹	۰/۰۷۸۹	۰/۰۹۴۴
	۰/۷۵	۰/۰۴۴۸	۰/۰۹۹۳	۰/۰۷۲۲	۰/۰۶۳۵
	۰/۹۹	۰/۰۵۷۴	۰/۱۸۱۳	۰/۰۳۰۸	۰/۰۴۵۴
$\beta_{۱۲۱}$	۰/۱	۰/۱۱۹۷	۰/۱۵۲۳	۰/۵۴۵۱	۰/۱۴۱۸
	۰/۲۵	۰/۰۵۴۱	۰/۱۴۹۶	۰/۳۳۴۲	۰/۰۸۸۹
	۰/۵	۰/۰۴۵۸	۰/۱۴۹۴	۱/۰۶۲۹	۰/۰۸۰۱
	۰/۷۵	۰/۰۲۶۱۲	۰/۰۵۰۲	۰/۰۴۷۳	۰/۰۵۹۵
	۰/۹۹	۰/۰۲۰۳	۰/۰۶۲۱	۰/۱۱۷۱	۰/۱۰۷۴
$\beta_{۲۲۰}$	۰/۱	۱/۵۰۷۶	۰/۰۵۵۹	۰/۳۱۵۲	۰/۱۳۳۶
	۰/۲۵	۰/۰۸۵۵	۰/۰۵۵۱	۰/۲۸۳۴	۰/۴۵۵۴
	۰/۵	۰/۰۶۱۳	۰/۰۵۲۹	۰/۱۱۲۰	۰/۱۲۷۹
	۰/۷۵	۰/۴۲۰۱	۰/۰۴۸۷	۰/۰۶۹۱	۰/۲۷۰۸
	۰/۹۹	۰/۰۵۶۱	۰/۰۲۰۶	۰/۰۴۳۱	۰/۰۴۱۱
$\beta_{۲۲۱}$	۰/۱	۰/۶۵۵۳	۰/۱۲۰۱	۰/۱۱۰۲	۱/۰۴۲۸
	۰/۲۵	۰/۱۶۵۸	۰/۱۱۶۸	۰/۰۸۸۹	۰/۱۰۳۱
	۰/۵	۰/۱۱۶۲	۰/۱۲۲۹	۰/۰۵۰۰	۰/۰۸۱۰
	۰/۷۵	۰/۰۳۷۱	۰/۰۴۹۸	۰/۰۳۳۱	۰/۱۲۷۴
	۰/۹۹	۰/۰۳۶۷	۰/۰۲۷۵	۰/۰۲۸۸	۱/۸۸۶۹
$\beta_{۳۲۰}$	۰/۱	۰/۲۷۲۱	۰/۰۷۵۶	۰/۰۹۵۴	۰/۰۶۶۰
	۰/۲۵	۱/۷۵۹۹	۰/۰۷۸۸	۰/۰۸۶۷	۰/۰۷۲۷
	۰/۵	۰/۱۴۱۱	۰/۰۷۹۵	۰/۱۵۵۱	۰/۱۳۶۲
	۰/۷۵	۰/۰۵۴۸۹	۰/۰۶۰۲	۰/۰۵۸۷	۰/۱۳۹۹
	۰/۹۹	۰/۰۳۸۴	۰/۰۵۹۲	۰/۰۴۴۲	۰/۰۳۵۸
$\beta_{۳۲۱}$	۰/۱	۰/۰۹۲۸	۰/۱۳۹۴	۰/۱۳۰۲	۰/۰۹۴۳
	۰/۲۵	۰/۱۲۳۱	۰/۱۳۵۱	۰/۱۵۳۳	۰/۰۶۴۳
	۰/۵	۰/۰۶۷۸	۰/۱۳۱۶	۰/۱۲۹۵	۰/۱۵۳۷
	۰/۷۵	۰/۰۳۹۳	۰/۲۳۱۴	۰/۰۴۶۲	۰/۰۷۰۱
	۰/۹۹	۰/۰۳۷۴	۱/۹۵۴۸	۰/۰۴۵۴	۰/۰۳۹۲



شکل ۱: نمودارهای خودهمبستگی (الف) و اثر (ب) نمونه‌های حاصل از توزیع‌های پسین ضرایب رگرسیونی β_{111} (ردیف اول)، β_{211} (ردیف دوم)، β_{311} (ردیف سوم)، β_{121} (ردیف چهارم)، β_{221} (ردیف پنجم) و β_{321} (ردیف ششم)

جدول ۳: کارایی نسبی براوردگر بسامدی به براوردگر بیزی برای ضرایب آمیختگی به‌ازای مقادیر مختلف این پارامتر.

α	حجم نمونه			
	۲۵	۵۰	۱۰۰	۲۰۰
۰/۱	۰/۶۱۸۹	۰/۸۲۶۴	۳/۴۱۴۳	۳/۲۱۲۲
۰/۲۵	۰/۰۶۹۷	۰/۴۲۴۰	۳/۱۰۹۴	۳/۸۳۳۹
۰/۵	۰/۰۰۸۶	۰/۰۰۴۶	۰/۰۰۳۳	۰/۰۰۲۹
۰/۷۵	۰/۳۱۶۸	۰/۱۵۰۱	۴/۱۰۴۵	۳/۱۸۲۳
۰/۹۹	۰/۱۷۸۰	۰/۳۸۲۷	۵/۰۶۷۴	۳/۵۹۷۶

۵ تحلیل داده‌های سرطان روده و معده

در این بخش نحوه کاربست دو رهیافت بسامدی و بیزی پیشنهادی در قالب یک مثال کاربردی شرح داده شده و کارایی آن مورد ارزیابی قرار گرفته است. برای این منظوراز مجموعه داده‌های مربوط به تعداد کل موارد بروز سرطان‌های روده بزرگ و معده در ۲۴ استان کشور طی سال‌های ۱۳۸۳ تا ۱۳۸۷ که در جدول ۴ ارائه شده‌اند، استفاده شده است. در این جدول ستون اضافه وزن نشان‌دهنده درصد افرادی است که شاخص BMI برای آن‌ها بیش‌تر از ۲۵ است. ستون کم تحرکی نیز درصد افرادی را نشان می‌دهد که دارای فعالیت فیزیکی کم‌تر از ۶۰۰ متر در هفته هستند. بر اساس آمارهای موجود تعداد موارد ابتلا به این دو نوع سرطان که از نظر علم پزشکی دارای رابطه معنی‌داری هستند، به شدت رو به افزایش است. عوامل مختلفی بر بروز این نوع سرطان‌ها موثر هستند که از آن جمله می‌توان به مصرف سیگار، اضافه وزن، کم‌تحرکی و میزان مصرف میوه و سبزیجات، اشاره کرد. در این مطالعه، هدف بررسی رابطه بین بروز سرطان‌های معده و روده بزرگ با اضافه وزن و کم‌تحرکی است. از این رو، متغیرهای مربوط به اضافه وزن و کم‌تحرکی به عنوان متغیرهای تبیینی در نظر گرفته شده‌اند. با توجه به مقادیر میانگین و واریانس متغیرهای پاسخ که در جدول ۵ ارائه شده اند، فرض نابرابری میانگین و واریانس مشاهدات پاسخ که یکی از مفروضات اصلی تحلیل رگرسیون پواسون معمول است، برقرار نیست و مشاهدات بسیار بیش پراکنده‌اند. از این رو، مدل رگرسیون پواسون

جدول ۴: تعداد کل موارد بروز سرطان‌های معده و روده بزرگ در سال‌های ۱۳۸۳ تا ۱۳۸۷ و عوامل خطر مورد مطالعه برای سرطان‌ها در سال ۱۳۸۷

ردیف	نام استان	سرطان معده	سرطان روده بزرگ	اضافه وزن	کم‌تحرکی
۱	اردبیل	۸۲۳	۶۱۷	۴۶/۲۸	۴۸/۸۲
۲	اصفهان	۹۶۳	۵۱۶	۴۷/۱۷	۳۸/۰۱
۳	ایلام	۱۱۰	۲۰۸	۳۵/۳۵	۳۱/۴۶
۴	بوشهر	۷۵	۴۹	۴۱/۵۱	۴۳/۱۲
۵	چهار محال و بختیاری	۲۵۸	۱۲۹	۳۹/۶۲	۳۸/۵۴
۶	خراسان جنوبی	۸۲	۱۷۴	۴۱/۰۲	۳۱/۸۵
۷	خراسان شمالی	۱۵۸	۴۱	۴۳/۲۵	۳۲/۲۴
۸	خوزستان	۵۷۴	۵۹۱	۴۶/۵۳	۴۵/۲۲
۹	زنجان	۴۱۱	۶۰۷	۳۵/۱۷	۵۰/۸۷
۱۰	سمنان	۲۳۰	۱۵۶	۴۰/۲۲	۵۰/۱۴
۱۱	سیستان و بلوچستان	۱۶۶	۱۵۸	۴۴/۸۳	۴۱/۰۷
۱۲	فارس	۸۷۷	۱۲۵	۴۶/۱۷	۴۶/۰۶
۱۳	قزوین	۳۶۲	۲۱۰	۴۵/۳۹	۳۴ /۴۰
۱۴	قم	۲۹۲	۲۰۶	۳۸/۵۲	۴۲/۸۰
۱۵	کردستان	۷۳۵	۱۹۷	۴۸/۶۷	۳۲/۳۵
۱۶	کرمان	۴۲۴	۳۲۱	۴۴/۵۱	۳۰/۲۷
۱۷	کرمانشاه	۵۱۶	۳۶۹	۴۵/۸۹	۳۸/۳۵
۱۸	کهگیلویه و بویراحمد	۲۰۹	۷۵	۴۵/۸۸	۲۹/۷۳
۱۹	گلستان	۵۲۰	۳۲۷	۵۳/۸۵	۲۱/۵۷
۲۰	لرستان	۵۰۶	۲۲۹	۴۴/۶۹	۲۴/۳۷
۲۱	مرکزی	۳۱۹	۲۲۳	۳۸/۵۰	۳۳/۵۹
۲۲	هرمزگان	۱۱	۱۲۲	۴۷/۳۱	۵۳/۷۵
۲۳	همدان	۴۶۶	۲۶۶	۲۹/۷۷	۳۱/۴۷
۲۴	یزد	۱۷۹	۲۱۵	۴۱/۰۱	۴۴/۲۶

دومتغیره آمیخته از طریق دو رهیافت بیزی و بسامدی به داده‌ها برازش داده شده و براورد پارامترها در جدول ۶ ارائه شده‌اند. برای مقایسه‌ی نتایج حاصل از برازش

جدول ۵: خلاصه آمارهای توصیفی متغیرهای پاسخ و تبیینی

متغیر	کمینه	چارک اول	میانه	میانگین	چارک سوم	بیشینه	انحراف معیار
سرطان معده	۷۵	۱۷۵/۸	۳۴۰/۷۵	۳۹۰/۲	۵۱۷	۹۶۳	۲۵۹/۵۵۹۵
سرطان روده	۴۱	۱۲۸	۲۰۷	۲۳۳/۳	۲۷۹/۸	۶۱۷	۱۵۹/۲۰۹۹
اضافه وزن	۲۹/۷۷	۴۰/۰۷	۴۴/۶	۴۲/۰۵	۴۶/۲	۵۳/۸۵	۵/۶۰۷۸۰
کم‌تحرکی	۲۱/۵۷	۳۱/۸۰	۳۸/۱۸	۳۸/۱۰	۴۴/۵۰	۵۳/۷۵	۸/۵۸۶۰۲

جدول ۶: برآورد پارامترهای مدل رگرسیون پواسون دومتغیره آمیخته

رهیافت	پارامتر					
	$\beta_{۱۱۱}$	$\beta_{۱۱۲}$	$\beta_{۲۱۱}$	$\beta_{۲۱۲}$	$\beta_{۳۱۱}$	$\beta_{۳۱۲}$
بیزی	۰/۵۰۸۸	۰/۴۶۱۹	۰/۴۷۱۸	۰/۵۲۰۵	۰/۴۵۲۹	۰/۵۲۰۹
بسامدی	۰/۷۴۶۴	۰/۸۲۸۵	۰/۷۹۲۰	۰/۸۰۰۷	۰/۷۸۴۵	۰/۷۸۸۵
	$\beta_{۱۲۱}$	$\beta_{۱۲۲}$	$\beta_{۲۲۱}$	$\beta_{۲۲۲}$	$\beta_{۳۲۱}$	$\beta_{۳۲۲}$
بیزی	۱/۶۸۱۶	۰/۴۲۳۷	۲/۲۵۷۴	۰/۵۱۱۸	−۰/۴۰۴۶	−۰/۶۴۰۰
بسامدی	−۱/۰۹۷۴	۰/۴۲۸۴	−۰/۴۹۰۷	۰/۱۸۹۵	۰/۵۰۲۱	−۰/۳۷۶۶

مدل رگرسیون پواسون دومتغیره آمیخته در دو رهیافت بیزی و بسامدی، از معیار ارزیابی متقابل $CV = \frac{۱}{n}\sum_{i=۱}^{n}(Y_i - \hat{Y}_{-i})^۲$ استفاده شده است، که درآن Y_i مقدار متغیر پاسخ iام و $\hat{Y}_{-i}$ مقدار پیش‌بینی شده این مشاهده بر اساس تمام مشاهدات به‌جز خود مشاهده است. مقدار آماره CV برای پیش‌بینی‌های حاصل از دو رهیافت بیزی و بسامدی در جدول ۷ ارائه شده‌اند. ملاحظه می‌شود که مقدار این آماره برای رهیافت بیزی پیشنهاد شده در این مقاله، به‌صورت قابل توجهی از مقدار متناظر برای رهیافت بسامدی کوچک‌تر است. با توجه به کوچکی حجم نمونه در این مثال، بار دیگر نتیجه حاصل از مطالعه شبیه‌سازی در بخش قبل مبنی بر این‌که مدل

بسامدی توسعه داده شده توسط برمودز و کارلیس (۲۰۱۲) نمی‌تواند برای مطالعاتی که در آن‌ها حجم نمونه بزرگ نیست به نتایج رضایت بخشی منجر شود، مورد تایید قرار می‌گیرد.

جدول ۷: مقدار آماره ارزیابی متقابل برای پیش‌بینی‌ها

رهیافت	سرطان معده	سرطان روده بزرگ
بیزی	۳/۴۰	۳/۹۸
بسامدی	۱۹/۷۲	۱۶/۹۲

۶ بحث و نتیجه‌گیری

استفاده از مدل رگرسیون پواسون آمیخته دومتغیره یکی از راه‌حل‌های ارائه شده جدید برای مواجهه با مشکلات بیش‌پراکنش و تورم مشاهدات در صفر، در تحلیل رگرسیونی داده‌های شمارشی است. با این مدل می‌توان همبستگی‌های ذاتی متغیرهای پاسخ را در فرایند مدل‌سازی لحاظ نمود و به مدل دقیق‌تری دست یافت. گر چه کاربست این مدل از هر دو دیدگاه بسامدی و بیزی به دلیل پیچیدگی‌های ذاتی مدل‌های آمیخته نیازمند استفاده از الگوریتم‌های تکراری بوده و با پیچیدگی‌های خاص خود همراه است، اما نتایج حاصل از شبیه‌سازی و نیز مثال کاربردی ارائه شده در این مقاله نشان می‌دهد که میزان کارایی رهیافت بیزی در براورد ضرایب رگرسیونی، دست کم به‌ازای اندازه نمونه‌های کوچک، به طور قابل ملاحظه‌ای از کارایی رهیافت بسامدی بیش‌تر است. به‌علاوه، استفاده از رهیافت بسامدی از طریق کاربست الگوریتم EM با مسائلی مانند تعیین مقادیر اولیه مناسب برای پارامترهای مدل مواجه است، که به دلیل پیچیدگی مدل و زیاد بودن تعداد پارامترهای آن کار ساده‌ای نیست. به طور کلی، تنها برای اندازه نمونه‌های بزرگ است که به دلیل بروز خواص مجانبی براوردگرهای ماکسیمم درستنمایی، می‌توان رهیافت بسامدی را در تحلیل رگرسیون پواسون آمیخته دومتغیره توصیه نمود.

تقدیر و تشکر

نویسندگان از پیشنهادهای ارزشمند داوران محترم که موجب ارتقای سطح کیفی مقاله شد، سپاس‌گزاری می‌کنند.

مراجع

Bermúdez, L. (2009), A Priori Ratemaking Using Bivariate Poisson Regression Models. *Insurance: Mathematics and Economics*, **44**, 135-141.

Bermúdez, L. and Karlis, D. (2011), Bayesian Multivariate Poisson Models for Insurance Ratemaking, *Insurance: Mathematics and Economics*, **48**, 226-236.

Bermúdez, L. and Karlis, D. (2012), Mixture of Bivariate Poisson Regression Models with an Application to Insurance Ratemaking, *Computational Statistics and Data Analysis*, **56**, 3988-3999.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B. (1994), The Distribution of the Likelihood Ratio for Mixtures of Densities from the One-parameter Exponential Family, *Annals of the Institute of Statistical Mathematics*, **46**, 373-388.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data Via the EM Algorithm, *Journal of Royal Statistical Society*, B, **39**, 1-38.

Früwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer Series in Statistics, Springer, New York.

Garay, A. M., Hashimoto, E. M., Ortega, E. M. M. and Lachos, V. H. (2011), On Estimation and Influence Diagnostics for Zero-inflated

Negative Binomial Regression Models, *Computational Statistics and Data Analysis*, **55**, 1304-1318.

Gardner, W., Mulvey E. P. and Shaw, E. C. (1995), Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson and Negative Binomial Models, *Psychological Bulletin*, **118**, 392-404.

Gelman, A., Jakulin, A., Grazia Pittau, M. and Su, Y. (2008), A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models, *The Annals of Applied Statistics*, **4**, 1360-1383.

Gelman, A. and Rubin, D. B. (1992), Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, **7**, 457-511.

George, E. L. and McCulloch, R. E. (1993), Variable Selection Via Gibbs Sampling, *Journal of American Statistical Association*, **88**, 881-890.

Gurmu, S. and Elder, J. (2008), A Bivariate Zero-inflated Count Data Regression Model with Unrestricted Correlation, *Economics Letters*, **100**, 245-248.

Karlis, D. and Ntzoufras, L. (2003), Analysis of Sports Data by Using Bivariate Poisson Models, *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**, 381-393.

Lambert, D. (1992), Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing, *Technometrics*, **34**, 1-14.

Marin, J. M., Mengersen, K. and Robert, C. (2005), *Bayesian Modelling and Inference on Mixtures of Distributions*, In Rao, C. and Dey, D., Editors, Handbook of Statistics, 25, Springer-Verlag, New York.

Raftery, A. E., Madigan, D. and Hoeting, J. (1997), Bayesian Model Averaging for Linear Regression Models, *Journal of the American Statistical Association*, **92**, 179-191.

Wang, P. (2003), A Bivariate Zero-inflated Negative Binomial Regression Model for Count Data with Excess Zeros, *Economics Letters*, **78**, 373-378.

Wang, P., Cockburn, I. M. and Puterman, M. L. (1998), Analysis of Patent Data: A Mixed Poisson Regression Model Approach, *Journal of Business and Economic Statistics* , **6**, 27-36.