

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 نادى فر، 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. به منظور ارزیابی مدل بیزی پیشنهادی، مطالعه‌ای شبیه‌سازی انجام شده و کارایی برآوردهای 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 موضوع بندی 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 دو جمله‌ای 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 دو جمله‌ای منفی در صفر متورم نیز توسط وانگ (۲۰۰۳) و گارای و همکاران (۲۰۱۱) مورد مطالعه قرار گرفته است. یکی از راه‌های مقابله با مشکل وجود صفرهای بیش از حد در مشاهدات پاسخ، استفاده از توزیع پواسون دو متغیره است. مدل‌های در صفر متورم در حالت دو متغیره توسط کارلیس و ان‌سوفراس (۲۰۰۳) و گورمو و والد (۲۰۰۸) مورد مطالعه قرار گرفته‌اند. اخیراً 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 دو متغیره را برای بررسی همبستگی ذاتی دو ادعا در صنعت بیمه مورد استفاده قرار دادند. یکی از مزیت‌های استفاده از توزیع‌های دو متغیره 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 دو متغیره را مطرح و نحوه برآورد 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 دو متغیره 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 به دلیل شکل پیچیده تابع درستنمایی مبتنی بر توزیع پواسون دو متغیره، 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 برآورد پارامترهای مدل در مقایسه با مدل بسامدی رقیب مورد ارزیابی قرار گرفته است.

در بخش ۲ مدل رگرسیون آمیخته متناهی پواسون دو متغیره و داده‌های ناقص و کامل به طور اجمالی معرفی و سپس رهیافت بسامدی برآورد پارامترهای مدل بر اساس روش ماکسیمم درستنمایی و الگوریتم EM، شرح داده شده است. در بخش ۳ رهیافت بیزی تحلیل رگرسیون آمیخته متناهی پواسون دو متغیره 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 دو متغیره 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 دو متغیره

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

۸۰. مدل رگرسیونی بیزی پواسون آمیخته دو متغیره

حالت چندمتغیره استفاده از این توزیع در عمل دشوار است (برمودز و کارلیس، ۲۰۱۱). یک راه حل معقول برای رفع مشکل بیش پراکنش، استفاده از مدل رگرسیون پواسون با متغیر پاسخ آمیخته است. توزیع پواسون دو متغیره به صورت

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

تعریف می شود. بر این اساس، مدل رگرسیون پواسون دو متغیره با متغیر پاسخ آمیخته متناهی را می توان به ازای $i = 1, \dots, n$ به صورت

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

نوشت، که در آن $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$ ، $\lambda_{\ell j}(\mathbf{x}_{\ell i}) = \exp(\mathbf{x}'_{\ell i} \boldsymbol{\beta}_{\ell j})$ ، $\ell = 1, 2, 3$ بردار متغیرهای تبیینی مربوط به لامین پارامتر $\mathbf{x}'_{\ell i} = (x_{1\ell i}, \dots, x_{p\ell i})$ ، $j = 1, \dots, g$ توزیع پواسون دو متغیره و $\boldsymbol{\beta}'_{\ell j} = (\beta_{1\ell j}, \dots, \beta_{p\ell j})$ بردار ضرایب رگرسیونی مربوط به لامین پارامتر توزیع پواسون دو متغیره و λ زامین زیرجامعه است. مجموعه پارامترهای مدل رگرسیون پواسون دو متغیره با متغیر پاسخ آمیخته متناهی g مؤلفه ای به صورت

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

و مشتمل بر $3gp + g$ پارامتر است. فرض کنید نمونه ای تصادفی به حجم n به صورت $\mathcal{D} = \{((y_{1i}, y_{2i}), \mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})\}_{i=1}^n$ موجود است. چون از نحوه انتساب مشاهدات به زیرجامعه های توزیع آمیخته اطلاعاتی در دست نیست، چنین داده هایی

را اصطلاحاً ناقص می نامند و تابع درستنمایی آن‌ها به صورت

$$\begin{aligned}
 L(\boldsymbol{\theta}, \boldsymbol{\alpha}) &= \prod_{i=1}^n \left\{ \sum_{j=1}^g \alpha_j BP(y_{1i}, y_{r_i} \mid \lambda_{1j}(\mathbf{x}_{1i}), \lambda_{r_j}(\mathbf{x}_{r_i}), \lambda_{r_j}(\mathbf{x}_{r_i})) \right\} \\
 &= \prod_{i=1}^n \left\{ \sum_{j=1}^g [\alpha_j e^{-(\lambda_{1j}(\mathbf{x}_{1i}) + \lambda_{r_j}(\mathbf{x}_{r_i}) + \lambda_{r_j}(\mathbf{x}_{r_i}))} \right. \\
 &\quad \times \frac{\lambda_{1j}^{y_{1i}}(\mathbf{x}_{1i}) \lambda_{r_j}^{y_{r_i}}(\mathbf{x}_{r_i})}{y_{1i}! y_{r_i}!} \\
 &\quad \times \left. \sum_{s_i=0}^{\min(y_{1i}, y_{r_i})} \binom{y_{1i}}{s_i} \binom{y_{r_i}}{s_i} s_i! \left(\frac{\lambda_{r_j}(\mathbf{x}_{r_i})}{\lambda_{1j}(\mathbf{x}_{1i}) \lambda_{r_j}(\mathbf{x}_{r_i})} \right)^{s_i} \right\},
 \end{aligned}$$

تعریف می شود. به کمک بردار متغیرهای گم شده $z_i = (z_{i1}, \dots, z_{ig})^T$ که نشان‌دهنده وضعیت انتساب هر مشاهده به مؤلفه‌های توزیع آمیخته است و به صورت

$$z_{ij} = \begin{cases} 1 & \text{اگر } y_i = (y_{1i}, y_{r_i}) \text{ به } z \text{ متعلق به } j \text{ باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

تعریف می شود، نسخه دیگری از داده‌ها که اصطلاحاً داده‌های کامل نامیده می شوند، ساخته می شود. چون هر مشاهده تنها می تواند 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 z_i دارای توزیع چندجمله‌ای $\mathcal{M}(\alpha_1, \dots, \alpha_g)$ هستند. از این رو، تابع درستنمایی داده‌های کامل به صورت

$$\begin{aligned}
 L_c(\boldsymbol{\theta}, \boldsymbol{\alpha}) &= P(\mathcal{D} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}) \\
 &= \prod_{i=1}^n \prod_{j=1}^g \{ [BP(y_{1i}, y_{r_i} \mid \beta_{1j}, \beta_{r_j}, \beta_{r_j}, \mathbf{x}_{1i}, \mathbf{x}_{r_i}, \mathbf{x}_{r_i})]^{z_{ij}} \alpha_j^{z_{ij}} \} \\
 &= \prod_{i=1}^n \left(\prod_{j=1}^g [e^{-(\exp(\mathbf{x}'_{1i} \beta_{1j}) + \exp(\mathbf{x}'_{r_i} \beta_{r_j}) + \exp(\mathbf{x}'_{r_i} \beta_{r_j}))} \right. \\
 &\quad \times \frac{(\exp(\mathbf{x}'_{1i} \beta_{1j}))^{y_{1i}} (\exp(\mathbf{x}'_{r_i} \beta_{r_j}))^{y_{r_i}}}{y_{1i}! y_{r_i}!} \\
 &\quad \times \sum_{s_i=0}^{\min(y_{1i}, y_{r_i})} \binom{y_{1i}}{s_i} \binom{y_{r_i}}{s_i} s_i! \\
 &\quad \times \left. \left(\frac{\exp(\mathbf{x}'_{r_i} \beta_{r_j})}{\exp(\mathbf{x}'_{1i} \beta_{1j}) \exp(\mathbf{x}'_{r_i} \beta_{r_j})} \right)^{s_i} \right]^{z_{ij}} \prod_{j=1}^g \alpha_j^{z_{ij}}, \quad (2)
 \end{aligned}$$

۸۲ مدل رگرسیون بیزی پواسون آمیخته دو متغیره

خواهد بود، که در آن $\mathcal{D} = \left\{ \left(\begin{matrix} y_{1i} \\ y_{2i} \end{matrix} \right), x_{1i}, x_{2i}, x_{3i} \right\}_{i=1}^n$ نشان‌دهنده مشاهدات و $\eta_j = \sum_{i=1}^n z_{ij}$ بیان‌گر تعداد مشاهدات مربوط به زیرجامعه j ام توزیع آمیخته است.

۱.۲ رهیافت بسامدی برازش مدل رگرسیون آمیخته پواسون دو متغیره

چون برآوردگرهای ماکسیمم درست‌نمایی پارامترهای مدل رگرسیون پواسون دو متغیره با متغیر پاسخ آمیخته متناهی فرم بسته ندارند، یکی از روش‌های رایج برای برآورد ماکسیمم درست‌نمایی پارامترهای این مدل، استفاده از الگوریتم EM است (دمپستر و همکاران، ۱۹۷۷). در این مدل بردار متغیرهای پنهان به صورت

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

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

$$L_c(\Phi) = \prod_{i=1}^n \prod_{j=1}^g \left(\alpha_j BP(y_{1i}, y_{2i} | \lambda_{1j}(x_{1i}), \lambda_{2j}(x_{2i}), \lambda_{3j}(x_{3i})) \right)^{z_{ij}},$$

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

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

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

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

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

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

خواهد بود. چون z_{ij} یک متغیر تصادفی برنولی است، وقتی i امین مشاهده به z_{ij} 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

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

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

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

(۲) مرحله E. در مرحله t ام مقادیر

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

را محاسبه کنید، که در آن‌ها

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

۸۴ مدل رگرسیونی بیزی پواسون آمیخته دو متغیره

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

(۳) مرحله M. براساس مقادیر پارامترها در مرحله t ام، برآورد پارامترها در مرحله $(t+1)$ ام را به صورت

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

محاسبه کنید، که در آن $\mathbf{b}_{3j} = [b_{1j}, \dots, b_{nj}]^T$ یک بردار $n \times 1$ بعدی و $\hat{\beta}(\mathbf{y}, \mathbf{x}, \mathbf{w})$ برآوردگر ماکسیمم درست‌نمایی موزون پارامترهای یک مدل پواسون دو متغیره با بردار پاسخ \mathbf{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 برآوردهای حاصل از الگوریتم EM هستند. برای بررسی همگرایی الگوریتم EM قواعد مختلفی پیشنهاد شده است. به عنوان نمونه می توان از قاعده ساده‌ای به صورت

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

استفاده کرد، که در آن ξ کمیت کوچکی است و آستانه همگرایی نامیده می شود (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. معمولاً توصیه می شود مقادیر اولیه مختلفی مورد آزمایش قرار گیرند تا بهترین نرخ همگرایی و برازش حاصل شود (دمپستر و همکاران، ۱۹۷۷).

به طور کلی، هنگامی که مدل پیچیده و تعداد پارامترها زیاد باشد، تعیین مقادیر اولیه مناسب برای پارامترها کار ساده‌ای نیست. از طرفی، چون خواص بهینگی برآوردگرهای 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

به طور کلی تحلیل بیزی مدل‌های آمیخته دارای پیچیدگی‌های زیاد و دشواری‌های خاص خود است. توزیع پسین در این مدل‌ها معمولاً پیچیده و فاقد فرم بسته است. به‌علاوه نمونه‌گیری مناسب از توزیع پسین اغلب امری دشوار و زمان‌بر است (مارین و همکاران، ۲۰۰۵). برآورد 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 دو متغیره با متغیر پاسخ آمیخته متناهی g مؤلفه‌ای، توزیع پیشین پارامترها را می‌توان با فرض استقلال پیشینی ضرایب رگرسیونی و ضرایب آمیختگی به صورت

$$\pi(\Phi) = \prod_{j=1}^g [\pi(\beta_{1j}, \beta_{2j}, \beta_{\tau j})] \pi(\alpha_1, \dots, \alpha_g), \quad (5)$$

نوشت. معمولاً برای ضرایب آمیختگی، $\alpha = (\alpha_1, \dots, \alpha_g)$ ، توزیع پیشین دیریکله به صورت

$$\begin{aligned} \pi(\alpha_1, \dots, \alpha_g) &= \text{Dirichlet}(c_1, \dots, c_g) \\ &= \frac{1}{B(c)} \prod_{j=1}^g \alpha_j^{c_j-1} \end{aligned} \quad (6)$$

در نظر گرفته می شود. این انتخاب در ادبیات مدل های آمیخته، یکی از انتخاب های مرسوم و کلاسیک است (فرورس-شناتر، ۲۰۰۶). از آن جا که توزیع دیریکله تعمیم یافته توزیع بتا است، بر اساس رابطه ای که بین توزیع بتا و $u(0, 1)$ وجود دارد و به منظور پرهیز از پیش داوری معمولاً ابرپارامترهای توزیع پیشین به صورت $c_1 = c_2 = \dots = c_g = c_0$ و برابر در نظر گرفته می شود. اگر $c_0 = 1$ باشد، آن گاه توزیع یکنواخت d -بعدی، $u(0, 1)^d$ ، حاصل می شود. البته فرض برابری ابرپارامترهای توزیع دیریکله زمانی مطلوب است که همه ضرایب آمیختگی با هم برابر باشند، به این معنی که تحلیل گر معتقد باشد، همه مؤلفه های توزیع آمیخته به یک اندازه در نمونه مشاهده شده سهم دارند. از این رو، باید تعلق مشاهدات به هر یک از زیرجوامع را در نظر گرفت. یک روش برای این منظور در حالت های یک و دو متغیره رسم نمودار مشاهدات و مقایسه فراوانی مشاهدات در مؤلفه های تشکیل دهنده توزیع آمیخته است. البته این کار در حالت های چندمتغیره دشوار است. به عنوان مثال، اگر در یک مدل آمیخته دو مؤلفه ای، فراوانی مشاهدات مربوط به مؤلفه اول از فراوانی مشاهدات مربوط به مؤلفه دوم کم تر (بیش تر) باشد، ابرپارامترهای توزیع دیریکله را باید به نحوی تعیین کرد که این توزیع چوله به راست (چپ) باشد.

بسیاری از محققان، از جمله گلמן و همکاران (۲۰۰۸)، استفاده از توزیع پیشین نرمال برای ضرایب مدل های رگرسیونی را در حالت کلی منطقی می دانند. بر این اساس، برای ضرایب رگرسیونی $(\beta_{1j}, \beta_{2j}, \beta_{3j})$ ، $j = 1, \dots, g$ ، توزیع نرمال چندمتغیره به صورت

$$\pi(\beta_{1j}, \beta_{2j}, \beta_{3j}) \sim N_{r_p}(\mu_j, \Sigma_j), \quad j = 1, \dots, g. \quad (V)$$

به عنوان 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، می توان به استنباط پسینی پیرامون پارامترها پرداخت.

۲.۳ توزیع پسین

با توجه به توزیع های پیشین (۷) و (۷)، توزیع پسین به صورت

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

به دست می آید. به دلیل پیچیدگی توزیع پسین، امکان محاسبه ثابت چگالی ساز عملاً وجود ندارد. از این رو، توزیع پسین فاقد شکل بسته و شناخته شده است. بر این اساس، می توان از زنجیر مارکف مونت کارلویی برای نمونه گیری از این توزیع پسین استفاده کرد. سپس نمونه های حاصل را برای هر گونه استنباط پسینی پیرامون پارامترها به کار گرفت.

۱.۲.۳ توزیع های پسین شرطی کامل

در این بخش نحوه کاربست الگوریتم گیبز برای نمونه گیری از توزیع پسین (۸) شرح داده می شود. لازمه استفاده از الگوریتم گیبز شناخت توزیع های پسین شرطی کامل است. بنابراین در این بخش توزیع های پسین شرطی کامل برای ضرایب آمیختگی، ضرایب رگرسیون و متغیرهای گم شده z_i محاسبه می شوند. بر اساس

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

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

محاسبه می شود. بنابراین، توزیع پسین شرطی کامل برای ضرایب آمیختگی، توزیع دیریکله به صورت

$$\pi(\alpha | \mathbf{Z}, \boldsymbol{\theta}, \mathcal{D}) = \text{Dirichlet}(c_1 + \eta_1, \dots, c_g + \eta_g),$$

است. به همین ترتیب توزیع پسین شرطی کامل ضرایب رگرسیونی $\boldsymbol{\theta} = \{\beta_{1j}, \beta_{2j}, \beta_{3j}\}_{j=1}^g$ به صورت

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

است، که در آن $\varphi_{3p}(\cdot, \cdot, \cdot)$ تابع چگالی توزیع نرمال $3p$ -بعدی است. توزیع پسین شرطی کامل متغیر نشان گر z_{ij} که وضعیت تعلق i امین مشاهده به i امین مؤلفه را مشخص می سازد، به صورت

$$P(z_{ij} | \boldsymbol{\theta}, \alpha, \mathcal{D}) = \frac{\alpha_j BP(y_{1i}, y_{2i} | \beta_{1j}, \beta_{2j}, \beta_{3j}, \mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})}{\sum_{j=1}^g \alpha_j BP(y_{1i}, y_{2i} | \beta_{1j}, \beta_{2j}, \beta_{3j}, \mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})} \quad (9)$$

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

$$\pi(z_i | \theta, \alpha, \mathcal{D}) = \mathcal{M}(\lambda, Pr(z_{i1} | \theta, \alpha, \mathcal{D}), \dots, Pr(z_{ig} | \theta, \alpha, \mathcal{D})),$$

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

$$\pi(z_i | \theta, \alpha, \mathcal{D}) = \mathcal{M}(\lambda, Pr(z_{i1} | \theta, \alpha, \mathcal{D}), \dots, Pr(z_{ig} | \theta, \alpha, \mathcal{D})),$$

$$\pi(\alpha | \mathbf{Z}, \theta, \mathcal{D}) = \text{Dirichlet}(c_1 + \eta_1, \dots, c_g + \eta_g),$$

$$\begin{aligned} \pi(\theta | \alpha, \mathbf{Z}, \mathcal{D}) &\propto \prod_{i=1}^n \left(\prod_{j=1}^g [e^{-(\exp(\mathbf{x}'_i \beta_{1j}) + \exp(\mathbf{x}'_i \beta_{2j}) + \exp(\mathbf{x}'_i \beta_{3j}))}] \right) \\ &\times \frac{(\exp(\mathbf{x}'_i \beta_{1j}))^{y_{1i}} (\exp(\mathbf{x}'_i \beta_{2j}))^{y_{2i}}}{y_{1i}! y_{2i}!} \\ &\times \sum_{s_i=0}^{\min(y_{1i}, y_{2i})} \binom{y_{1i}}{s_i} \binom{y_{2i}}{s_i} s_i! \\ &\times \left(\frac{\exp(\mathbf{x}'_i \beta_{3j})}{\exp(\mathbf{x}'_i \beta_{1j}) \exp(\mathbf{x}'_i \beta_{2j})} \right)^{s_i}]^{z_{ij}} \\ &\times \prod_{j=1}^g \varphi_{rp}(\beta_{1j}, \beta_{2j}, \beta_{3j}; \mu_j, \Sigma_j). \end{aligned}$$

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

(۱) مقدار اولیه $\alpha^{(0)}$ را برای α و $\theta^{(0)} = \{\beta_{1j}^{(0)}, \beta_{2j}^{(0)}, \beta_{3j}^{(0)}\}_{j=1}^g$ را برای θ در نظر بگیرید.

^۳ Gibbs sampling

^۴ Metropolis-Hastings

۹۰مدل رگرسیونی بیزی پواسون آمیخته دو متغیره

(۲) در مرحله $(t + 1)$ ام بر اساس مقادیر حاصل از مرحله t ام، α ، Z ، θ را به صورت زیر به روز کنید:

$$\pi(Z_i^{(t+1)} | \theta^{(t)}, \alpha^{(t)}, \mathcal{D}) = \mathcal{M}(1, P(z_i^{(t+1)} | \theta^{(t)}, \alpha^{(t)}, \mathcal{D}), \dots, P(z_{ig}^{(t+1)} | \theta^{(t)}, \alpha^{(t)}, \mathcal{D})),$$

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

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

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

در مرحله ۲ الگوریتم، برای نمونه گیری از توزیع پسین شرطی کامل $\{\theta | Z, \alpha, \mathcal{D}\}$ به دلیل بسته نبودن شکل توزیع، از الگوریتم متروپولیس - هستینگس برای نمونه گیری از این توزیع استفاده می شود.

۴ مطالعه شبیه سازی

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

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

در نظر گرفته شده‌اند. به علاوه، به منظور 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} ، $i = 1, \dots, n$ از توزیع نرمال $N(\mu, \sigma^2)$ شبیه‌سازی شده‌اند. بر این اساس، با توجه به نرمال بودن ترکیب خطی متغیرهای نرمال و رابطه $\lambda(x) = \exp(x'\beta)$ ، متغیر تصادفی $\lambda(X)$ دارای توزیع لگ نرمال $LN(\mu, \sigma^2)$ است. سپس مقادیر پاسخ $y_i = (y_{1i}, y_{2i})$ از طریق رابطه (۲) شبیه‌سازی شده است. در نهایت با فرض موجود بودن مشاهدات $\{y_i, x_i\}_{i=1}^n$ ضرایب رگرسیونی با رهیافت بسامدی (برمودز و کارلیس، ۲۰۱۲) و هم‌چنین رهیافت بیزی پیشنهاد شده در این مقاله، برآورد و میزان کارایی نسبی دو رهیافت در برآورد ضرایب رگرسیونی و پارامتر آمیختگی بر اساس معیار کمترین توان‌های دوم خطا^۵ (MSE) به صورت

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

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

به دست آمده‌اند. در تمامی محاسبات به منظور 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، معمولاً از نمودارهای اثر و خودهمبستگی برای تشخیص همگرایی استفاده می‌شود. به علاوه، برخی معیارهای عددی برای اطمینان بیش‌تر از همگرایی زنجیرهای مارکف، توسط محققانی مانند گلن 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 زنجیره‌های مارکف تولید شده از توزیع‌های پسین ضرایب رگرسیونی را نشان می‌دهند. البته با توجه به پیچیدگی مدل، همگرایی زنجیره‌های مارکف خصوصا برای نمونه‌های بزرگ، زمان‌بر است.

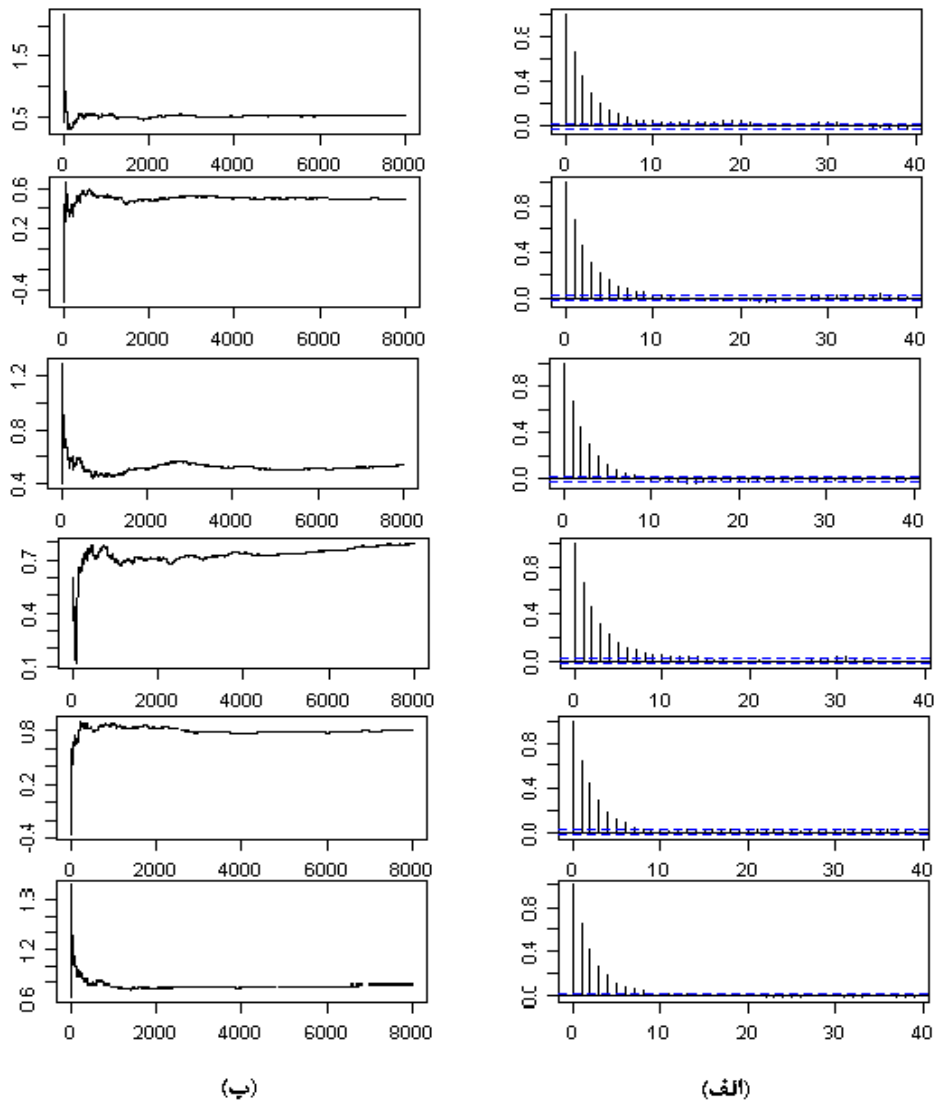
همان‌طور که ملاحظه می‌شود، میزان کارایی رهیافت بیزی در برآورد ضرایب رگرسیونی به‌ازای تمام مقادیر α و تمام اندازه نمونه‌های در نظر گرفته شده، از کارایی رهیافت بسامدی بیش‌تر است. برتری کارایی برآوردگر حاصل از مدل بیزی به برآوردگر بسامدی ماکسیمم درست‌نمایی به‌ویژه برای مقادیر ضریب آمیختگی نزدیک به ۰/۵ کاملاً قابل توجه است. البته چنان‌چه انتظار می‌رود، کارایی برآوردگرهای ماکسیمم درست‌نمایی بر اساس خواص مجانبی این نوع برآوردگرها با افزایش اندازه نمونه بهبود می‌یابد. در مورد پارامتر ضریب آمیختگی، α ، که در تحلیل رگرسیونی مستقیماً مورد علاقه نیست، وضعیت تا حدودی متفاوت است. به‌گونه‌ای که رهیافت بیزی تنها به‌ازای مقادیر متوسط α برای تمام اندازه‌های نمونه‌ای از رهیافت بسامدی کارا تر است (ردیف سوم جدول ۲). برای مقادیر α نزدیک به ۰ یا ۱ برتری کارایی رهیافت بیزی به بسامدی تنها برای نمونه‌های کوچک {۲۵، ۵۰} صادق است و برآوردگر 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 عملاً 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 برآوردگر بسامدی نسبت به برآوردگر بیزی برای ضرایب رگرسیون به‌ازای مقادیر مختلف ضرایب آمیختگی مولفه اول

حجم نمونه				α	$\beta_{\varepsilon_{jk}}$
۲۰۰	۱۰۰	۵۰	۲۵		
۰/۰۸۷۷	۰/۱۳۷۹	۰/۰۴۹۸	۰/۰۳۳۴	۰/۱	
۰/۴۳۴۳	۰/۱۱۷۷	۰/۰۵۱۷	۰/۰۲۶۶	۰/۲۵	
۰/۰۱۹۳	۰/۰۷۹۸	۰/۰۵۱۷	۰/۰۱۹۳	۰/۵	β_{110}
۰/۱۶۹۹	۰/۱۱۸۰	۰/۲۱۶۵	۰/۱۶۹۹	۰/۷۵	
۰/۰۷۹۷	۰/۱۶۱۷	۰/۰۶۵۶	۰/۰۷۹۷	۰/۹۹	
۰/۳۷۸۸	۰/۰۵۶۹	۰/۶۴۹۵	۰/۰۲۲۵	۰/۱	
۰/۰۸۷۳	۱/۷۱۸۴	۰/۵۸۳۲	۰/۰۲۰۱	۰/۲۵	
۰/۶۵۸۶	۰/۱۷۰۹	۰/۶۱۹۷	۰/۰۵۹۴	۰/۵	β_{111}
۰/۲۲۷۶	۰/۱۱۵۶	۰/۰۹۲۷	۰/۳۳۰۸	۰/۷۵	
۰/۲۱۸۶	۰/۱۰۳۲	۱/۱۳۲۳	۰/۱۷۹۱	۰/۹۹	
۰/۱۰۵۰	۰/۲۴۰۱	۰/۰۴۰۸	۰/۱۲۳۲	۰/۱	
۰/۲۱۸۳	۰/۲۴۵۸	۰/۰۴۲۶۷	۰/۰۷۹۹	۰/۲۵	
۰/۱۳۷۱	۰/۰۶۸۴	۰/۰۴۱۳	۰/۰۵۴۴	۰/۵	β_{210}
۰/۹۴۱۲	۰/۲۲۶۹	۰/۰۸۷۹	۰/۱۲۷۶	۰/۷۵	
۰/۴۳۳۱	۰/۶۹۰۹	۰/۰۷۵۶	۰/۰۷۶۸	۰/۹۹	
۰/۰۷۱۰	۰/۰۵۰۷	۰/۱۱۶۶	۰/۰۵۶۲	۰/۱	
۰/۲۹۹۹	۰/۱۰۲۵	۰/۱۲۶۶	۰/۰۷۳۷	۰/۲۵	
۰/۴۲۸۱	۰/۰۸۲۳	۰/۱۱۸۲	۰/۰۵۳۳	۰/۵	β_{211}
۰/۹۴۷۷	۰/۰۸۹۴	۰/۲۵۲۱	۰/۱۱۹۸	۰/۷۵	
۰/۱۴۳۶	۰/۱۲۶۴	۰/۲۳۰۳	۰/۲۰۲۹	۰/۹۹	
۰/۱۳۵۵	۱/۴۲۲۵	۰/۰۴۳۹	۰/۰۲۹۷	۰/۱	
۰/۰۸۳۵	۰/۸۴۲۴	۰/۰۴۴۶	۰/۰۳۱۲	۰/۲۵	
۰/۱۸۶۱	۰/۰۲۳۹	۰/۰۴۳۸	۰/۰۲۴۴	۰/۵	β_{210}
۰/۱۳۵۲	۰/۱۰۳۳	۰/۰۶۲۰	۰/۰۲۷۴	۰/۷۵	
۰/۱۶۶۴	۰/۲۴۳۴	۰/۰۶۴۰	۰/۰۱۷۲	۰/۹۹	
۰/۲۰۱۱	۰/۱۷۸۳	۰/۳۸۱۱	۰/۶۳۸۶	۰/۱	
۰/۱۰۵۴	۰/۱۲۲۹	۰/۳۵۷۷	۰/۹۰۳۵	۰/۲۵	
۰/۱۳۷۴	۰/۰۷۸۰	۰/۳۷۳۲	۰/۰۴۶۲	۰/۵	β_{211}
۰/۱۰۰۴	۰/۱۵۶۹	۰/۱۷۶۷	۰/۰۴۷۲	۰/۷۵	
۰/۱۳۷۹	۰/۸۵۹۹	۰/۳۱۵۳	۰/۰۵۲۱	۰/۹۹	

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

حجم نمونه				α	$\beta_{\epsilon_{jk}}$
۲۰۰	۱۰۰	۵۰	۲۵		
۰/۱۵۳۱	۰/۱۷۵۱	۰/۱۲۹۰	۱/۲۰۶۷	۰/۱	β_{1r_0}
۰/۰۶۹۳	۰/۱۵۱۶	۰/۱۲۲۰	۰/۲۳۴۰	۰/۲۵	
۰/۰۹۴۴	۰/۰۷۸۹	۰/۱۱۹۹	۰/۱۵۲۹	۰/۵	
۰/۰۶۳۵	۰/۰۷۲۲	۰/۰۹۹۳	۰/۰۴۴۸	۰/۷۵	
۰/۰۴۵۴	۰/۰۳۰۸	۰/۱۸۱۳	۰/۰۵۷۴	۰/۹۹	
۰/۱۴۱۸	۰/۵۴۵۱	۰/۱۵۲۳	۰/۱۱۹۷	۰/۱	β_{1r_1}
۰/۰۸۸۹	۰/۳۳۴۲	۰/۱۴۹۶	۰/۰۵۴۱	۰/۲۵	
۰/۰۸۰۱	۱/۰۶۲۹	۰/۱۴۹۴	۰/۰۴۵۸	۰/۵	
۰/۰۵۹۵	۰/۰۴۷۳	۰/۰۵۰۲	۰/۰۲۶۱۲	۰/۷۵	
۰/۱۰۷۴	۰/۱۱۷۱	۰/۰۶۲۱	۰/۰۲۰۳	۰/۹۹	
۰/۱۳۳۶	۰/۳۱۵۲	۰/۰۵۵۹	۱/۵۰۷۶	۰/۱	β_{2r_0}
۰/۴۵۵۴	۰/۲۸۳۴	۰/۰۵۵۱	۰/۰۸۵۵	۰/۲۵	
۰/۱۲۷۹	۰/۱۱۲۰	۰/۰۵۲۹	۰/۰۶۱۳	۰/۵	
۰/۲۷۰۸	۰/۰۶۹۱	۰/۰۴۸۷	۰/۴۲۰۱	۰/۷۵	
۰/۰۴۱۱	۰/۰۴۳۱	۰/۰۲۰۶	۰/۰۵۶۱	۰/۹۹	
۱/۰۴۲۸	۰/۱۱۰۲	۰/۱۲۰۱	۰/۶۵۵۳	۰/۱	β_{2r_1}
۰/۱۰۳۱	۰/۰۸۸۹	۰/۱۱۶۸	۰/۱۶۵۸	۰/۲۵	
۰/۰۸۱۰	۰/۰۵۰۰	۰/۱۲۲۹	۰/۱۱۶۲	۰/۵	
۰/۱۲۷۴	۰/۰۳۳۱	۰/۰۴۹۸	۰/۰۳۷۱	۰/۷۵	
۱/۸۸۶۹	۰/۰۲۸۸	۰/۰۲۷۵	۰/۰۳۶۷	۰/۹۹	
۰/۰۶۶۰	۰/۰۹۵۴	۰/۰۷۵۶	۰/۲۷۲۱	۰/۱	β_{3r_0}
۰/۰۷۲۷	۰/۰۸۶۷	۰/۰۷۸۸	۱/۷۵۹۹	۰/۲۵	
۰/۱۳۶۲	۰/۱۵۵۱	۰/۰۷۹۵	۰/۱۴۱۱	۰/۵	
۰/۱۳۹۹	۰/۰۵۸۷	۰/۰۶۰۲	۰/۰۵۴۸۹	۰/۷۵	
۰/۰۳۵۸	۰/۰۴۴۲	۰/۰۵۹۲	۰/۰۳۸۴	۰/۹۹	
۰/۰۹۴۳	۰/۱۳۰۲	۰/۱۳۹۴	۰/۰۹۲۸	۰/۱	β_{3r_1}
۰/۰۶۴۳	۰/۱۵۳۳	۰/۱۳۵۱	۰/۱۲۳۱	۰/۲۵	
۰/۱۵۳۷	۰/۱۲۹۵	۰/۱۳۱۶	۰/۰۶۷۸	۰/۵	
۰/۰۷۰۱	۰/۰۴۶۲	۰/۲۳۱۴	۰/۰۳۹۳	۰/۷۵	
۰/۰۳۹۲	۰/۰۴۵۴	۱/۹۵۴۸	۰/۰۳۷۴	۰/۹۹	



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_{411} (ردیف چهارم)، β_{511} (ردیف پنجم) و β_{611} (ردیف ششم)

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

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

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 منظور از 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 شده‌اند، 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	کم تحرکی
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

۹۸ مدل رگرسیونی بیزی پواسون آمیخته دو متغیره

دو متغیره آمیخته از طریق دو رهیافت بیزی و بسامدی به داده‌ها برازش داده شده و برآورد 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
کم تحرکی	۲۱/۵۷	۳۱/۸۰	۳۸/۱۸	۳۸/۱۰	۴۴/۵۰	۵۳/۷۵	۸/۵۸۶۰۲

جدول ۶: برآورد پارامترهای مدل رگرسیون پواسون دو متغیره آمیخته

رهیافت	پارامتر					
	β_{111}	β_{112}	β_{211}	β_{212}	β_{311}	β_{312}
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
رهیافت	پارامتر					
	β_{121}	β_{122}	β_{221}	β_{222}	β_{321}	β_{322}
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 دو متغیره آمیخته در دو رهیافت بیزی و بسامدی، از معیار ارزیابی متقابل $CV = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_{-i})^2$ استفاده شده است، که در آن مقدار متغیر پاسخ Y_i و \hat{Y}_{-i} مقدار پیش‌بینی شده این مشاهده بر اساس تمام مشاهدات به جز 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، به صورت 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 این که مدل

بسامدی توسعه داده شده توسط برمودز و کارلیس (۲۰۱۲) نمی تواند برای مطالعاتی که در آن ها حجم نمونه بزرگ نیست به نتایج رضایت بخشی منجر شود، مورد تایید قرار می گیرد.

جدول ۷: مقدار آماره ارزیابی متقابل برای پیش بینی ها

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

۶ بحث و نتیجه گیری

استفاده از مدل رگرسیون پواسون آمیخته دو متغیره یکی از راه حل های ارائه شده جدید برای مواجهه با مشکلات بیش پراکنش و تورم مشاهدات در صفر، در تحلیل رگرسیونی داده های شمارشی است. با این مدل می توان همبستگی های ذاتی متغیرهای پاسخ را در فرایند مدل سازی لحاظ نمود و به مدل دقیق تری دست یافت. گرچه کاربست این مدل از هر دو دیدگاه بسامدی و بیزی به دلیل پیچیدگی های ذاتی مدل های آمیخته نیازمند استفاده از الگوریتم های تکراری بوده و با پیچیدگی های خاص خود همراه است، اما نتایج حاصل از شبیه سازی و نیز مثال کاربردی ارائه شده در این مقاله نشان می دهد که میزان کارایی رهیافت بیزی در برآورد ضرایب رگرسیونی، دست کم به ازای اندازه نمونه های کوچک، به طور قابل ملاحظه ای از کارایی رهیافت بسامدی بیش تر است. به علاوه، استفاده از رهیافت بسامدی از طریق کاربست الگوریتم EM با مسائلی مانند تعیین مقادیر اولیه مناسب برای پارامترهای مدل مواجهه است، که به دلیل پیچیدگی مدل و زیاد بودن تعداد پارامترهای آن کار ساده ای نیست. به طور کلی، تنها برای اندازه نمونه های بزرگ است که به دلیل بروز خواص مجانبی برآوردگرهای ماکسیمم درست نمایی، می توان رهیافت بسامدی را در تحلیل رگرسیون پواسون آمیخته دو متغیره 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.

مدل رگرسیونی بیزی پواسون آمیخته دو متغیره ۱۰۲

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.