مجله علوم آماری، پاییز و زمستان ۱۳۹۶
جلد ۱۱، شماره ۲، ص ۲۴۱ – ۲۶۲
DOI: 10.18869/acadpub.jss.11.2.241

مدل‌بندی داده‌های شمارشی تحت تأثیر بیش‌پراکنش با مدل رگرسیون پواسون- بیرنبام ساندرز

رضا پورموسی و نرجس گیلانی
دانشگاه شهید باهنر کرمان، دانشکده ریاضی و کامپیوتر، بخش آمار
تاریخ دریافت: ۱۳۹۵/۱۰/۶ تاریخ آخرین بازنگری: ۱۳۹۶/۹/۲۵

چکیده: در این مقاله ابتدا به معرفی مدل‌های رگرسیون پواسون آمیخته پرداخته و در ادامه به معرفی یک مدل جدید به نام رگرسیون پواسون-بیرنبام ساندرز با هدف لحاظ کردن مسئله بیش‌پراکنش در مدل‌بندی داده‌های شمارشی پرداخته می‌شود. از آن‌جا که توزیع بیرنبام ساندرز آمیخته‌ای از دو توزیع گاوسی وارون تعمیم‌یافته است، لذا می‌توان مدل معرفی شده دو پارامتری را تعمیمی بر مدل‌های قبلی دانست که علاوه بر داشتن یک پارامتر کمتر نسبت به مدل رگرسیون پواسون گاوسی وارون تعمیم‌یافته، دارای شکل بسته در تابع جرم احتمال حاشیه‌ای و گشتاورهای مربوطه است. برای برآورد پارامترهای این مدل از الگوریتم EM استفاده و در نهایت کارایی این مدل نسبت به مدل‌های موجود با استفاده از مطالعه شبیه‌سازی شده و یک مثال واقعی نشان داده شده است.
واژه‌های کلیدی: وزیع بیرنبام ساندرز، الگوریتم EM، داده‌های شمارشی، مدل‌های رگرسیون پواسون، بیش‌پراکنش، مدل رگرسیون پواسون آمیخته.

۱ مقدمه

رگرسیون پواسون نوعی از مدل‌های خطی تعمیم‌یافته است که برای تحلیل داده‌های حاصل از شمارش به کار می‌رود. در این مدل متغیر پاسخ نشانگر تعداد اتفاقات در یک بازه زمانی است. در مدل‌های مبتنی بر توزیع نرمال، میانگین و واریانس معمولاً از هم مستقل هستند، در حالی‌که در بعضی از توزیع‌ها مانند

آدرس الکترونیکی نویسنده مسئول مقاله: رضا پورموسی، pourm@uk.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62E15، 62J12

پواسون و دوجمله‌ای، میانگین و واریانس به هم وابسته‌اند. اگر مجموعه‌ای از مشاهدات y_i از توزیع پواسون پیروی کنند، انتظار آن است که واریانس آن‌ها برابر با میانگین باشد.

تحلیل آماری داده‌های شمارشی سابقه طولانی و غنی دارد و توزیع پواسون نقش کلیدی در تحلیل این نوع داده‌ها را ایفا می‌کند (جانسون و همکاران، ۱۹۹۲؛ تیلور و کارلین، ۱۹۹۴). یکی از ویژگی‌های کلیدی توزیع پواسون برابری میانگین و واریانس آن است:

$$E(Y) = Var(Y) = \mu.$$

چارچوب اصلی در مدل‌بندی داده‌های شمارشی، استفاده از مدل رگرسیون پواسون است (پاتیل، ۱۹۷۰)، اما به دلیل محدودیتی که در ساختار این توزیع وجود دارد (برابری واریانس با میانگین) محققان معمولاً به دنبال تعمیم مدل‌هایی برای مدل‌بندی داده‌های شمارشی بیش‌پراکنده هستند.

هایند (۱۹۸۲) و کارلیس و زیکالاکی (۲۰۰۵) به بررسی کارایی مدل رگرسیون پواسون، مسئله بیش‌پراکنش[1] و کم‌پراکنش[2] در داده‌های شمارشی پرداخته‌اند. هم‌چنین سائز کستیلو و کوند سانچز (۲۰۱۳) مدل‌های رگرسیون پواسون آمیخته و بعضی از ویژگی‌های آن‌ها را مورد مطالعه قرار داده‌اند. بیش‌پراکنش یک مسئله مهم در داده‌های شمارشی است که در ادبیات مدل‌های خطی تعمیم‌یافته راه‌حل‌های مختلفی برای آن پیشنهاد شده است. توزیع‌های جایگزین برای تحلیل داده‌های شمارشی توزیع‌های پواسون آمیخته هستند که به دلیل وجود پارامترهای بیشتر از انعطاف‌پذیری بالاتری نسبت به مدل پواسون برخوردار هستند.

وجود بیش‌پراکنش انگیزه‌ای است که به دنبال مدل‌های بهتری برای برازش این نوع داده‌ها باشیم. مدل‌های رگرسیون پواسون- گاما، پواسون- گاوسی وارون، پواسون- گاوسی وارون تعمیم‌یافته مثال‌هایی از این نتایج هستند (گاردنر و همکاران، ۱۹۹۵؛ استین و جوریتز، ۲۰۰۷؛ رایگبی و همکاران، ۲۰۰۸).

مسئله بیش‌پراکنش زمانی رخ می‌دهد که پارامتر توزیع پواسون به تنهایی قادر به توصیف رخدادهای شمارشی نیست. در حالت کلی بیش‌پراکنش دارای دو منبع ناهمگنی جامعه و وجود بیش از حد صفر است. هدف این مقاله ارائه یک مدل رگرسیون پواسون آمیخته[3] برای افزایش کارایی مدل در داده‌های بیش‌پراکنش است. برای این منظور در بخش‌های ۲ و ۳ مدل رگرسیون پواسون، توزیع پواسون آمیخته

[1] Overdispersion
[2] Underdispersion
[3] Mixed Poisson regression models

و خواص آن‌ها معرفی شده و مدل‌های رگرسیون پواسون- گاما[۴]، پواسون- گاوسی وارون[۵] و پواسون- گاوسی وارون تعمیم‌یافته[۶] که از این خانواده نتیجه شده‌اند به همراه ویژگی‌های آن‌ها تعریف شده‌اند. در بخش ۴ مدل جدیدی از رگرسیون پواسون آمیخته به نام پواسون- بیرنبام ساندرز[۷] معرفی و خواص آن مطرح می‌شود. سپس نشان داده می‌شود مدل معرفی شده می‌تواند برای مدل‌بندی داده‌های بیش‌پراکنش مورد استفاده قرار گیرد. در بخش ۵ برای برآورد پارامترهای مدل از الگوریتم EM استفاده و در نهایت در بخش ۶ کارایی این مدل در برازش و پیش‌بینی نسبت به مدل‌های موجود با استفاده از یک مطالعه شبیه‌سازی شده و یک مثال با داده‌های واقعی نمایش داده شده است.

۲ مدل رگرسیون پواسون و تعمیم‌های آن

فرض کنید در n مشاهده مستقل i امین مشاهده با $(y_i, \mathbf{x}_i)$ نمایش داده شود، که در آن y_i نشان دهنده تعداد وقوع یک پیشامد و $\mathbf{x}_i$ بردار مستقل خطی است.

تعریف ۱: (مدل رگرسیون پواسون) فرض کنید Y_i دارای توزیع پواسون با میانگین μ_i باشد، آنگاه:

$$f_{Y_i}(y_i|x_i) = \frac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!}, \quad \mu_i > ۰, \quad y_i = ۰, ۱, \ldots$$

و چون میانگین توزیع پواسون، نشان‌دهنده تعداد مورد انتظار پیشامدها و غیرمنفی است، از مدل لگاریتم میانگین به جای یک مدل خطی استفاده می‌شود، یعنی:

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}. \qquad (۱)$$

حال با نمایی کردن رابطه (۱) یک مدل ضربی برای میانگین به‌صورت

$$E(Y_i|\mathbf{x}_i) = \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}), \qquad (۲)$$

به دست می‌آید، که به آن مدل رگرسیون پواسون گویند. در رابطه (۲)، $\boldsymbol{\beta}$ بردار پارامترهای رگرسیون، $\mathbf{x}_i$

[4] Poisson-Gamma
[5] Poisson- inverse Gaussian
[6] Poisson-generalized inverse Gaussian
[7] Poisson- Birnbaum Saunders

بردار متغیرهای تبیینی مستقل خطی به‌صورت

$$\boldsymbol{\beta}^T = [\beta_0, \ldots, \beta_p], \qquad \mathbf{x}_i^T = [1, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip}].$$

هستند. به منظور مدل‌بندی داده‌های شمارشی معمولاً فرض

$$Var(Y) = \phi E(Y) = \phi\mu,$$

به عنوان تناسب میانگین و واریانس در نظر گرفته می‌شود، که در آن اگر $\phi = 1$ باشد، آنگاه واریانس برابر با میانگین است و اگر $\phi > 1$ باشد، آنگاه نسبت به توزیع پواسون بیش‌پراکنش وجود دارد.

۳ توزیع‌های پواسون آمیخته

تعریف ۲: (توزیع پواسون آمیخته) فرض کنید با توجه به متغیر تصادفی λ که با تابع احتمال حاشیه‌ای $g_\lambda(\lambda)$ روی R^+ تعریف می‌شود، Y دارای توزیع پواسون با میانگین $\mu\lambda$، ($\mu > 0$) باشد، در این صورت توزیع حاشیه‌ای Y، توزیع پواسون آمیخته با تابع چگالی

$$P_g(y|\phi) = \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!} \quad dG(\lambda|\phi),$$

است، که در آن $G(\lambda|\phi)$ تابع توزیع متغیر تصادفی λ است.

گزاره ۱: اگر متغیر تصادفی Y دارای توزیع پواسون آمیخته باشد، آنگاه میانگین و واریانس آن عبارتند از

$$E(Y) = E(E(Y|\lambda)) = \mu E(\lambda),$$

$$Var(Y) = \mu\, E(\lambda) + \mu^2\, Var(\lambda).$$

طبق رابطه (۲) چون پارامتر μ نقش مدل رگرسیون پواسون را به عهده دارد، لذا با استفاده از گزاره ۱، در توزیع‌های پواسون آمیخته، پارامترهای توزیع آمیزنده را طوری در نظر می‌گیرند که $E(\lambda) = 1$ شود. در

این صورت

$$E(Y) = \mu, \tag{۳}$$

$$Var(Y) = \mu + \mu^۲ Var(\lambda). \tag{۴}$$

با توجه به رابطه (۴) واریانس توزیع پواسون آمیخته همواره بزرگتر از واریانس توزیع پواسون است. معمولا توزیع‌های متعددی را به عنوان توزیع آمیزنده مورد آزمایش قرار داده‌اند، معروف‌ترین آن‌ها که در ادامه از آن‌ها استفاده شده، توزیع‌های گاما($\Gamma(\alpha,\beta)$)، گاوسی وارون ($IG(\alpha,\beta)$ و گاوسی وارون تعمیم‌یافته ($GIG(\nu,\chi,\psi)$ هستند که تابع چگالی آن‌ها به ترتیب عبارتند از

$$f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)}\ y^{\alpha-۱}\ e^{-\beta y},\ \ y>۰,$$

$$f_Y(y) = \sqrt{\frac{\beta}{۲\pi y^۳}}\ \exp\{-\frac{\beta(y-\alpha)^۲}{۲\alpha^۲ y}\},\ \ y>۰,$$

$$f_Y(y) = \frac{(\frac{\psi}{\chi})^{\frac{\nu}{۲}}}{۲K_\nu(\sqrt{\psi\chi})} y^{\nu-۱}\exp(\frac{-۱}{۲}(\chi y^{-۱}+\psi y)),\ \ y>۰.$$

که در آن $K_\nu(\cdot)$ تابع بسل تعمیم یافته نوع سوم است (گرادشتین و ریزهیک، ۲۰۰۰).

با انتخاب توزیع‌های آمیزنده متفاوت، می‌توان توزیع‌های پواسون آمیخته جدیدی را تعریف نمود، امّا تعداد زیادی از آن‌ها به دلیل نداشتن شکل بسته کاربرد چندانی در مدل‌های پواسون آمیخته ندارند. در ادامه چند مدل معروف از این خانواده که با توزیع‌های آمیزنده گاما، گاوسی وارون و گاوسی وارون تعمیم‌یافته تولید شده‌اند، معرفی می‌شوند.

تعریف ۳: (مدل رگرسیون پواسون-گاما، $PO-Gamma$) فرض کنید Y به شرط متغیر تصادفی مشاهده نشده λ دارای توزیع پواسون با میانگین $\mu\lambda$ باشد، که در آن $\mu = \exp(\mathbf{x}^T\boldsymbol{\beta})$ و $\lambda \sim \Gamma(\frac{۱}{\sigma},\frac{۱}{\sigma})$. در این صورت تابع جرم احتمال حاشیه‌ای متغیر تصادفی Y (که همان توزیع دو جمله‌ای منفی است) به صورت

$$P(Y=y) = \frac{\Gamma(y+\frac{۱}{\sigma})}{\Gamma(y+۱)\ \Gamma(\frac{۱}{\sigma})}\ (\frac{\sigma\mu}{۱+\sigma\mu})^y\ (\frac{۱}{۱+\sigma\mu})^{\frac{۱}{\sigma}},\quad y=۰,۱,\ldots$$

با میانگین و واریانس

$$E(Y) = \mu, \quad Var(Y) = \mu(۱ + \sigma\mu)$$

است. محققان بسیاری مانند میاوی (۱۹۹۴) و گاردنر و همکاران (۱۹۹۵) به مقایسه عملکرد مدل‌های رگرسیون پواسون و رگرسیون پواسون-گاما پرداخته‌اند.

تعریف ۴: (مدل رگرسیون پواسون- گاوسی وارون، $PO - IG$) فرض کنید توزیع Y به شرط متغیر تصادفی مشاهده نشده λ پواسون با میانگین $\mu\lambda$ باشد، که در آن $\mu = \exp(\mathbf{x}^T\boldsymbol{\beta})$ و $\lambda \sim IG(۱, \frac{۱}{\sigma})$. در این صورت تابع جرم احتمال حاشیه‌ای Y به‌صورت

$$P(Y = y) = (\frac{۲\theta}{\pi})^{\frac{۱}{۲}} \ \frac{\mu^y}{y!} \ e^{\frac{۱}{\sigma}} \ \frac{K_{y-\frac{۱}{۲}}(\theta)}{(\theta\sigma)^y}, \quad y = ۰, ۱, \ldots$$

است، که در آن $\theta^۲ = \frac{۱}{\sigma^۲} + \frac{۲\mu}{\sigma}$ و میانگین و واریانس آن عبارتند از

$$E(Y) = \mu, \qquad Var(Y) = \mu(۱ + \sigma\mu).$$

استین و جوریتز (۲۰۰۷) در مقاله خود به بررسی مدل‌های خطی با توزیع پواسون- گاوسی وارون پرداخته‌اند.

تعریف ۵: (مدل رگرسیون پواسون- گاوسی وارون تعمیم‌یافته، $PO - GIG$) فرض کنید توزیع Y به شرط متغیر تصادفی مشاهده نشده λ پواسون با میانگین $\mu\lambda$ باشد، که در آن $\mu = \exp(\mathbf{x}^T\boldsymbol{\beta})$ و $\lambda \sim GIG(\nu, \frac{۱}{c\sigma}, \frac{c}{\sigma})$. در این صورت تابع جرم احتمال حاشیه‌ای Y به‌صورت

$$P(Y = y) = (\frac{\mu}{c})^y \ \frac{۱}{(\theta\sigma)^{y+\nu} y!} \ \frac{K_{\nu+y}(\theta)}{K_\nu(\frac{۱}{\sigma})}, \quad y = ۰, ۱, \ldots$$

است، که در آن

$$\theta^۲ = \frac{۱}{\sigma^۲} + \frac{۲\mu}{c\sigma}, \quad c = R_\nu(\frac{۱}{\sigma}) = \frac{K_{\nu+۱}(\frac{۱}{\sigma})}{K_{\nu(\frac{۱}{\sigma})}}$$

و میانگین و واریانس آن عبارتند از

$$E(Y)=\mu, \quad Var(Y)=\mu(۱+\mu(\frac{۲\sigma(\nu+۱)}{c}+\frac{۱}{c^{۲}}-۱)).$$

رایگبی و همکاران (۲۰۰۸) به معرفی ویژگی‌های توزیع پواسون- گاوسی وارون تعمیم‌یافته و چارچوبی برای مدل‌بندی داده‌های شمارشی پرداخته‌اند.

۴ معرفی مدل رگرسیون پواسون- بیرنبام ساندرز

در این بخش مدلی از خانواده رگرسیون پواسون آمیخته به نام پواسون-بیرنبام ساندرز معرفی می‌شود، که در آن λ دارای توزیع آمیزنده بیرنبام ساندرز فرض شده است. این ایده از ارتباط توزیع گاوسی وارون تعمیم‌یافته و توزیع بیرنبام ساندرز به دست آمده است. از مزیت‌های این مدل می‌توان به شکل بسته تابع احتمال آن و داشتن یک پارامتر کمتر نسبت به بعضی مدل‌های معرفی شده قبلی، اشاره کرد.

با توجه به این‌که توزیع بیرنبام ساندرز ترکیب خطی از دو توزیع گاوسی وارون تعمیم‌یافته است، به نظر می‌رسد که انتخاب توزیع بیرنبام ساندرز به عنوان توزیع آمیزنده در ساخت مدل‌های رگرسیون پواسون آمیخته به دلیل داشتن شکل بسته و هم‌چنین داشتن واریانس بزرگتر از میانگین، می‌تواند برای لحاظ کردن مسئله بیش‌پراکنش در داده‌های شمارشی مورد استفاده قرار گیرد. یک شبیه‌سازی با استفاده از بسته‌های نرم‌افزاری bs و $gamlss$ در محیط R به منظور ارزیابی مدل پیشنهادی و مقایسه نظری آن با مدل‌های مرسوم استفاده شده است. معیار ارزیابی عملکرد مدل پیشنهاد شده، درصد قدرت تشخیص صحیح مدل است، یعنی از سری داده‌های شبیه‌سازی شده و آلوده کردن آن‌ها با مسئله بیش‌پراکنش، در چند درصد از مواقع مدل پیشنهادی بهتر از سایر مدل‌های نامزد، عمل کرده است؟ هم‌چنین برای مقایسه نیکویی برازش مدل پیشنهادی و مدل‌های نامزد از ملاک MSE استفاده شده است.

اخیراً هاشیموتوا و همکاران (۲۰۱۳) یک مدل پواسون- بیرنبام ساندرز به منظور مدل‌بندی میزان بهبودی بقا و با این فرض که تعداد پیشامدها از توزیع پواسون و زمان پیشامدها از توزیع بیرنبام ساندرز پیروی می‌کنند، معرفی کرده‌اند. ایده ارائه شده توسط آن‌ها کاملاً متفاوت با ایده پیشنهادی در این مقاله است.

۱.۴ توزیع بیرنبام ساندرز و خواص آن

بسیاری از فلزات به کار رفته در پل‌ها، خودروها، هواپیماها و غیره در اثر اعمال بار به دفعات مکرر تحت تنش قرارگرفته و پس از یک دوره کار زیاد ترک خورده و کم کم می‌شکنند. این اتفاق معمولاً بدون اطلاع قبلی و غیر قابل رویت رخ می‌دهد. توزیع بیرنبام ساندرز که برای نخستین بار توسط بیرنبام و ساندرز (۱۹۶۹) معرفی شد به طور گسترده‌ای برای مدل‌بندی زمان شکست در آمار به کار برده می‌شود. زمان شکست همان زمان شروع ترک در فلزات تا رسیدن به شکست نهایی تلقی می‌شود. محققان بسیاری مانند دیازگارسیا و لیوا(۲۰۰۵)، فرام و لی (۲۰۰۶) و لیمونت و همکاران (۲۰۰۷) به بسط و تعمیم این توزیع پرداخته‌اند. هم‌چنین کاندو و همکاران (۲۰۱۰) توزیع بیرنبام ساندرز دو بعدی که دارای توزیع‌های حاشیه‌ای بیرنبام ساندرز تک بعدی است را معرفی کرده‌اند. این توزیع از دیدگاه بیزی ابتدا توسط آچکار (۱۹۹۳) مورد بررسی قرارگرفت، هم‌چنین نیل و رابرت (۲۰۰۸)، ژو و تانگ (۲۰۱۱) و وانگ و همکاران (۲۰۱۶) این توزیع را از دیدگاه بیزی مورد استنباط آماری قرار داده‌اند.

تعریف ۶: متغیر تصادفی W دارای توزیع بیرنبام ساندرز $BS(\alpha,\beta)$ است، اگر تابع توزیع تجمعی آن به‌صورت

$$F_{BS}(w;\alpha,\beta)=\Phi[\frac{۱}{\alpha}\{(\frac{w}{\beta})^{\frac{۱}{۲}}-(\frac{\beta}{w})^{\frac{۱}{۲}}\}],\quad w,\alpha,\beta>۰,$$

باشد، که در آن $\Phi(۰)$ تابع توزیع تجمعی نرمال استاندارد و α و β به ترتیب پارامترهای شکل و مقیاس هستند. در این صورت تابع چگالی احتمال w عبارت است از

$$f_{BS}(w;\alpha,\beta)=\frac{۱}{۲\sqrt{۲\pi}\alpha\beta}[(\frac{\beta}{w})^{\frac{۱}{۲}}+(\frac{\beta}{w})^{\frac{۳}{۲}}]\exp[-\frac{۱}{۲\alpha^{۲}}(\frac{w}{\beta}+\frac{\beta}{w}-۲)]. \qquad (۵)$$

گزاره ۲: امید ریاضی و واریانس توزیع بیرنبام ساندرز عبارتند از

$$E(W)=\beta(۱+\frac{۱}{۲}\alpha^{۲}),\qquad Var(W)=(\alpha\beta)^{۲}(۱+\frac{۵}{۴}\alpha^{۲}).$$

یکی از خواص مورد توجه در توزیع بیرنبام ساندرز این است که می‌توان آن را به صورت آمیخته‌ای از دو توزیع گاوسی وارون تعمیم‌یافته مشخص کرد که در گزاره ۳ به آن پرداخته شده است.

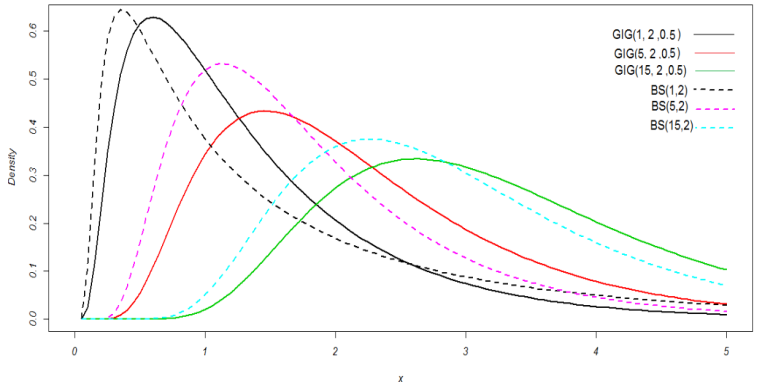
گزاره ۳: (کاندو و همکاران، ۲۰۱۰) اگر متغیر تصادفی W دارای توزیع $BS(\alpha,\beta)$ باشد، آنگاه W را می‌توان به‌صورت آمیخته‌ای از دو متغیر تصادفی گاوسی وارون تعمیم‌یافته مستقل از هم به‌صورت

$$W \overset{d}{=} \begin{cases} X_1 & \text{با احتمال } \frac{1}{2} \\ X_2 & \text{با احتمال } \frac{1}{2} \end{cases}$$

نوشت، که در آن $X_1 \sim GIG(-\frac{1}{2}, \frac{\beta}{\alpha^2}, \frac{1}{\beta\alpha^2})$ و $X_2 \sim GIG(\frac{1}{2}, \frac{\beta}{\alpha^2}, \frac{1}{\beta\alpha^2})$. در این صورت

$$f_{BS}(w;\alpha,\beta) = \frac{1}{2} f_{GIG}(w;\frac{1}{2},\frac{\beta}{\alpha^2},\frac{1}{\beta\alpha^2}) + \frac{1}{2} f_{GIG}(w;-\frac{1}{2},\frac{\beta}{\alpha^2},\frac{1}{\beta\alpha^2}).$$

در شکل ۱ نمودار توزیع‌های گاوسی وارون تعمیم‌یافته و بیرنبام ساندرز برای چند ترکیب مختلف از پارامترها مشخص شده است.



شکل ۱. نمودار تابع چگالی توزیع‌های گاوسی وارون تعمیم‌یافته و بیرنبام ساندرز برای چند ترکیب مختلف از پارامترها

۲.۴ مدل رگرسیون پواسون-بیرنبام ساندرز

در این بخش مدل جدید رگرسیون پواسون- بیرنبام ساندرز به منظور افزایش کارایی مدل رگرسیون پواسون توسط توزیع آمیزنده بیرنبام ساندرز به عنوان ترکیب خطی از دو توزیع گاوسی وارون تعمیم‌یافته برای مدل‌بندی داده‌های بیش‌پراکنش معرفی می‌شود.

قضیه ۱: فرض کنید Y به شرط متغیر تصادفی مشاهده نشده λ، دارای توزیع پواسون با میانگین و واریانس $\mu\lambda$ باشد که در آن $\mu = \exp(\mathbf{x}^T\boldsymbol{\beta})$ و λ دارای توزیع بیرنبام ساندرز $BS(\alpha, \beta)$ است. در این صورت تابع چگالی حاشیه‌ای Y به‌صورت

$$f_Y(y) = (\frac{\mu\beta}{\alpha^2\delta})^y \frac{e^{\frac{1}{\alpha^2}}}{\sqrt{2\pi}y!\alpha}[\alpha^{-1}\delta^{\frac{-1}{2}}K_{y+\frac{1}{2}}(\delta) + \alpha\delta^{\frac{1}{2}}K_{y-\frac{1}{2}}(\delta)],\ y = 0, 1, \ldots \quad (۶)$$

است، که در آن $\delta^2 = \frac{1}{\alpha^4} + \frac{2\mu\beta}{\alpha^2}$.

برهان: با توجه به رابطه (۵) می‌توان نوشت:

$$\begin{aligned} f_Y(y) &= \int_0^\infty f(y|\lambda)\, g(\lambda)\, d\lambda \\ &= \int_0^\infty \frac{e^{-\mu\lambda}(\mu\lambda)^y}{y!}\, \frac{1}{2\sqrt{2\pi}\alpha\beta}\, ((\frac{\beta}{\lambda})^{\frac{1}{2}} + (\frac{\beta}{\lambda})^{\frac{3}{2}})\, \exp\{\frac{-1}{2\alpha^2}(\frac{\lambda}{\beta} + \frac{\beta}{\lambda} - 2)\} \\ &= \frac{\mu^y\, e^{\frac{1}{\alpha^2}}}{2\sqrt{2\pi}\,\alpha\beta\, y!}\, [\beta^{\frac{1}{2}} \int_0^\infty \lambda^{y-\frac{1}{2}}\, e^{\frac{-1}{2}[(\frac{1}{\beta\alpha^2}+2\mu)\lambda + \frac{\beta}{\alpha^2}\lambda^{-1}]}\, d\lambda \\ &+ \beta^{\frac{3}{2}} \int_0^\infty \lambda^{y-\frac{3}{2}}\, e^{\frac{-1}{2}[(\frac{1}{\beta\alpha^2}+2\mu)\lambda + \frac{\beta}{\alpha^2}\,\lambda^{-1}]}\, d\lambda] \end{aligned}$$

انتگرال‌های موجود در رابطه فوق بر اساس خواص تابع بسل تعمیم‌یافته نوع سوم به‌صورت

$$\int_0^\infty \lambda^{y-\frac{1}{2}} e^{\frac{-1}{2}[(\frac{1}{\beta\alpha^2}+2\mu)\lambda + \frac{\beta}{\alpha^2}\lambda^{-1}]} d\lambda = 2(\frac{\alpha^2\delta}{\beta})^{-y-\frac{1}{2}}\ K_{y+\frac{1}{2}}(\delta)$$

به دست می‌آیند. هم‌چنین

$$\int_0^\infty \lambda^{y-\frac{3}{2}} e^{\frac{-1}{2}[(\frac{1}{\beta\alpha^2}+2\mu)\lambda + \frac{\beta}{\alpha^2}\,\lambda^{-1}]}\, d\lambda = 2(\frac{\alpha^2\delta}{\beta})^{-y+\frac{1}{2}}\ K_{y-\frac{1}{2}}(\delta),$$

که در آن‌ها $\delta^2 = \frac{1}{\alpha^4} + \frac{2\mu\beta}{\alpha^2}$. بنابراین

$$f_Y(y) = \frac{\mu^y\, e^{\frac{1}{\alpha^2}}}{2\sqrt{2\pi}\,\alpha\beta\, y!}\, [\frac{2\beta^{\frac{1}{2}}\, K_{y+\frac{1}{2}}(\delta)}{(\frac{\alpha^2\delta}{\beta})^{y+\frac{1}{2}}} + \frac{2\beta^{\frac{3}{2}}\, K_{y-\frac{1}{2}}(\delta)}{(\frac{\alpha^2\delta}{\beta})^{y-\frac{1}{2}}}],$$

که از آن رابطه (۶) حاصل می‌شود.

در قضیه ۱ برای اعمال فرض $E(\lambda) = ۱$، طبق رابطه (۳) بایستی پارامترهای α و β در توزیع بیرنبام ساندرز را به‌صورت $\alpha = \sigma^{\frac{۱}{۲}}$ و $\beta = \frac{۲}{\sigma+۲}$ در نظر گرفت.

قضیه ۲: مدل پواسون-بیرنبام ساندرز ارائه شده در رابطه (۶)، دارای ویژگی (۴) است.

برهان: طبق روابط (۳) و (۴) به سادگی می‌توان نشان داد:

$$E(Y) = \mu,$$
$$Var(Y) = \mu + \mu^{۲}(\frac{\sigma(۴+۵\sigma)}{(\sigma+۲)^{۲}}).$$

بنابراین این مدل ویژگی (۴) را دارا است.

۵ برآورد ماکسیمم درست‌نمایی مدل پواسون-بیرنبام ساندرز

به‌دست آوردن برآورد ماکسیمم درست‌نمایی پارامترهای مدل رگرسیون پواسون-بیرنبام ساندرز مستلزم حل معادلات غیر خطی با روش‌های عددی تکراری است. الگوریتم امید-ماکسیمم[۸] EM یک روش عمومی برای محاسبه برآورد ماکسیمم درست‌نمایی در یک مسئله با داده‌های گم‌شده یا ناکامل است. الگوریتم EM توسط دمپستر و همکاران (۱۹۷۷) ارائه شده است. کارلیس (۲۰۰۵) و زی و وی (۲۰۰۸) الگوریتم EM برای توزیع‌های پواسون آمیخته را مورد بررسی قرار دادند.

الگوریتم EM معمولاً برای مواردی که تابع درست‌نمایی شکل پیچیده‌ای دارد و برآورد بیشینه درست‌نمایی پارامترها با روش مستقیم کار غیر ممکنی است، مورد استفاده قرار می‌گیرد. در مدل رگرسیون پواسون بیرنبام ساندرز، فرض کنید $\theta = (\beta^T, \sigma)^T$ بردار پارامترهای نامعلوم، Y_{obs} نشان دهنده داده‌های مشاهده شده $\{Y_i;\ i = ۱, \ldots, n\}$ و $\{\lambda_i;\ i = ۱, \ldots, n\}$ دنباله‌ای از وزن‌ها باشند که به عنوان داده‌های گمشده Y_{mis} رفتار می‌کنند. به‌علاوه $Y_c = (Y_{obs}, Y_{mis})$ اشاره به مجموعه داده‌های کامل دارد.

زمانی که ماکسیمم کردن $\log L(Y_{obs}|\theta)$ نسبت به θ آسان نیست، الگوریتم EM، آن را به وسیله تکرار ماکسیمم‌سازی $E(\log L(Y_c|\theta))$، ماکسیمم می‌کند. الگوریتم d EM یک لگوریتم تکرار شونده است که در آن هر تکرار شامل دو مرحله است. فرض کنید $\theta^{(۰)}$ مقادیر اولیه الگوریتم EM باشند.

[8] EM algorithm

۱. **گام E،** مرحله گرفتن امید ریاضی: امید ریاضی لگاریتم تابع درست‌نمایی داده‌های کامل به شرط داده‌های مشاهده شده و مقدار پارامتر جاری، یعنی $E(\log\ L(Y_c|\theta)|y_{obs},\theta^{(k)}) = Q(\theta|\theta^{(k)})$ محاسبه می‌شود.

۲. **گام** M، مرحله ماکسیمم‌سازی: مقدار $\theta^{(k+1)}$ به گونه‌ای انتخاب می‌شود که تابع $Q(\theta|\theta^{(k)})$ برای هر θ ماکسیمم شود، به عبارت دیگر $\theta^{(k+1)} = Arg\max_\theta(Q(\theta|\theta^{(k)})$.

گام‌های E و M تا زمانی که دنباله به همگرایی برسد تکرار می‌شوند، به عبارت دیگر تا زمانی که $\log L(Y_c|\theta^{k+1}) - \log L(Y_c|\theta^k)$ به کمترین مقدار دلخواه برسد. به طور کلی الگوریتم **EM** به نقطه اولیه حساس است و یک انتخاب نامناسب به عنوان نقطه اولیه موجب کاهش سرعت همگرایی الگوریتم **EM** می‌شود. برای انجام گام E نیاز به محاسبه امید ریاضی توابعی از متغیر تصادفی شرطی $\lambda|Y=y$ است، که برای این منظور از ویژگی خانواده سری‌های توانی[9] استفاده می‌شود.

تعریف ۷: یک توزیع گسسته متعلق به خانواده توزیع‌های سری توانی نامیده می‌شود، اگر تابع جرم احتمال آن را بتوان به‌صورت

$$P(y|\lambda) = \frac{a_y\lambda^y}{A(\lambda)},$$

نشان داد، که در آن $a_y > 0$ و $A(\lambda)$ تابعی از λ است و به y بستگی ندارد. بسیاری از توزیع‌های گسسته شناخته شده مانند توزیع پواسون و دوجمله‌ای متعلق به این خانواده هستند. اگر پارامتر λ متغیر تصادفی باشد، یک سری توانی آمیخته با تابع احتمال زیر حاصل می‌شود.

$$P_g(y|\phi) = \int_0^\infty \frac{a_y\lambda^y}{A(\lambda)}g(\lambda|\phi)d\phi$$

لم ۱: (ساپاتیناس، ۱۹۹۵) اگر $Y|\lambda$ از توزیع‌های گسسته توانی پیروی کند و λ دارای تابع احتمال حاشیه‌ای $g(\lambda)$ باشد، آن‌گاه

$$E(\lambda^r|Y=y) = \frac{P_g(y+r|\phi)a_y}{P_g(y|\phi)a_{y+r}}.$$

[9] Power series family

قضیه ۳: فرض کنید Y به شرط λ دارای توزیع پواسون با میانگین $\mu\lambda$ و $\lambda \sim BS(\alpha, \beta)$ باشد، آنگاه

الف: تابع چگالی λ به شرط مشاهده y به صورت

$$f(\lambda|y) = c\,[\beta^{\frac{1}{2}}\ \lambda^{y-\frac{1}{2}}\ e^{\frac{-1}{2}((\frac{1}{\beta\alpha^2}+2\mu)\lambda+\frac{\beta}{\lambda\alpha^2})} + \beta^{\frac{3}{2}}\ \lambda^{y-\frac{3}{2}}\ e^{\frac{-1}{2}((\frac{1}{\beta\alpha^2}+2\mu)\lambda+\frac{\beta}{\lambda\alpha^2})}],$$

است، که در آن

$$c = \frac{1}{2\beta}\,(\frac{\delta\alpha^2}{\beta})^y\ [(\delta\alpha^2)^{\frac{-1}{2}}\ K_{y+\frac{1}{2}}(\delta)\ + (\delta\alpha^2)^{\frac{1}{2}}\ K_{y-\frac{1}{2}}(\delta)]^{-1}.$$

ب: گشتاور rام متغیر تصادفی λ به شرط مشاهده y برابر است با:

$$E(\lambda^r|Y=y) = (\frac{\beta}{\alpha^2\delta})^r[\frac{\alpha^{-1}\delta^{\frac{-1}{2}}K_{y+r+\frac{1}{2}}(\delta) + \alpha\delta^{\frac{1}{2}}K_{y+r-\frac{1}{2}}(\delta)}{\alpha^{-1}\delta^{\frac{-1}{2}}\ K_{y+\frac{1}{2}}(\delta) + \alpha\delta^{\frac{1}{2}}\ K_{y-\frac{1}{2}}(\delta)}]. \qquad (۷)$$

برهان: الف: با توجه به این‌که

$$f(\lambda|y) = \frac{f(y,\lambda)}{f(y)} = kf(y|\lambda)g(\lambda), \quad \int_0^\infty f(\lambda|y)d\lambda = 1$$

که در آن k تابعی از y و مستقل از λ است. بنابراین

$$\begin{aligned}
\int_0^\infty f(y|\lambda)d\lambda &= k\int_0^\infty f(y|\lambda)g(\lambda)d\lambda \\
&= k\int_0^\infty \frac{e^{-\mu\lambda}(\mu\lambda)^y}{y!}\frac{1}{2\sqrt{2\pi}\alpha\beta}[(\frac{\beta}{\lambda})^{\frac{1}{2}} \\
&\quad +(\frac{\beta}{\lambda})^{\frac{3}{2}}]\exp[-\frac{1}{2\alpha^2}(\frac{\lambda}{\beta}+\frac{\beta}{\lambda}-2)]d\lambda \\
&= \frac{k\mu^y e^{\frac{1}{\alpha^2}}}{2\sqrt{2\pi}\alpha\beta y!}\int_0^\infty e^{-\mu\lambda}\lambda^y[(\frac{\beta}{\lambda})^{\frac{1}{2}} \\
&\quad +(\frac{\beta}{\lambda})^{\frac{3}{2}}]\exp[-\frac{1}{2\alpha^2}(\frac{\lambda}{\beta}+\frac{\beta}{\lambda})]d\lambda \\
&= k'[\beta^{\frac{1}{2}}\int_0^\infty \lambda^{y-\frac{1}{2}}e^{\frac{-1}{2}((\frac{1}{\beta\alpha^2}+2\mu)\lambda+\frac{\beta}{\lambda\alpha^2})}d\lambda
\end{aligned}$$

$$+\beta^{\frac{۳}{۲}} \int_{۰}^{\infty} \lambda^{y-\frac{۳}{۲}} e^{\frac{-۱}{۲}((\frac{۱}{\beta\alpha^{۲}}+۲\mu)\lambda+\frac{\beta}{\lambda\alpha^{۲}})} d\lambda] = ۱$$

با توجه به خواص تابع بسل تعمیم‌یافته نوع سوم می‌توان نوشت:

$$c = k' = \frac{۱}{۲\beta}(\frac{\delta\alpha^{۲}}{\beta})^{y}[(\delta\alpha^{۲})^{\frac{-۱}{۲}} K_{y+\frac{۱}{۲}}(\delta) + (\delta\alpha^{۲})^{\frac{۱}{۲}} K_{y-\frac{۱}{۲}}(\delta)]^{-۱}$$

بنابراین

$$f(\lambda|y) = c[\beta^{\frac{۱}{۲}} \lambda^{y-\frac{۱}{۲}} e^{\frac{-۱}{۲}((\frac{۱}{\beta\alpha^{۲}}+۲\mu)\lambda+\frac{\beta}{\lambda\alpha^{۲}})} + \beta^{\frac{۳}{۲}} \lambda^{y-\frac{۳}{۲}} e^{\frac{-۱}{۲}((\frac{۱}{\beta\alpha^{۲}}+۲\mu)\lambda+\frac{\beta}{\lambda\alpha^{۲}})}].$$

برای اثبات (ب) می‌توان از قضیه ۳ یا ویژگی خانواده سری‌های توانی استفاده کرد. با توجه به این‌که توزیع پواسون متعلق به خانواده توزیع‌های سری توانی با $A(\lambda) = e^{\mu\lambda}, a_y = \frac{\mu^y}{y!}$ است، با استفاده از لم ۱ اثبات به سادگی انجام می‌شود.

رابطه (۷) برای برآورد پارامترهای مدل رگرسیون پواسون- بیرنبام ساندرز به کمک الگوریتم EM مورد استفاده قرار می‌گیرد. تابع درست‌نمایی یک نمونه از توزیع پواسون- بیرنبام ساندرز با پارامترهای $\theta = (\beta^T, \sigma)^T$ به‌صورت

$$\log L_Y(\theta) = \sum_{i=۱}^{n} \log f_{y_i}(y_i;\theta).$$

است. تابع درست‌نمایی داده‌های کامل به‌صورت

$$\begin{aligned}
\log L_{Y,\lambda}(\theta) &= \log \prod_{i=۱}^{n} f_{Y_i|\lambda_i}(y_i;\theta) f_{\lambda_i}(\lambda_i,\theta) \\
&= \sum_{i=۱}^{n} \log f_{Y_i|\lambda_i}(y_i;\theta) + \sum_{i=۱}^{n} \log f_{\lambda_i}(\lambda_i;\theta) \\
&= L^{(۱)}_{Y_n|\lambda_n}(\theta) + L^{(۲)}_{\lambda_n}(\theta),
\end{aligned}$$

است، که در آن $Y_n = (Y_1, \ldots, Y_n)$ و $\lambda_n = (\lambda_1, \ldots, \lambda_n)$ است و هم‌چنین

$$\begin{aligned}
L^{(1)}_{Y_n|\lambda_n}(\theta) &= \sum_{i=1}^{n} \log f_{Y_i|\lambda_i}(y_i;\theta) \\
&= -\sum_{i=1}^{n} \exp(x_i^T\beta)\lambda_i + \sum_{i=1}^{n} y_i\ (x_i^T\beta) \\
&+ \sum_{i=1}^{n} y_i \log(\lambda_i) - \sum_{i=1}^{n} \log y_i! \\
L^{(2)}_{\lambda_n}(\theta) &= -n\log(4\sqrt{2}\pi) + n\log(\sigma+2) - \frac{n}{2}\log\sigma + \frac{n}{2}\log 2 \\
&- \frac{3n}{2}\log(\sigma+2) - \frac{3}{2}\sum_{i=1}^{n}\log\lambda_i + \sum_{i=1}^{n}\log((\sigma+2)\lambda_i + 2) \\
&- \sum_{i=1}^{n}\frac{\lambda_i(\sigma+2)}{4\sigma} - \sum_{i=1}^{n}\frac{1}{\sigma(\sigma+2)\lambda_i} + \frac{n}{\sigma}
\end{aligned}$$

این روابط در الگوریتم EM مورد استفاده قرار می‌گیرند.

۶ مثال‌های کاربردی

در این بخش ابتدا به شبیه‌سازی از مدل پواسون- بیرنبام ساندرز معرفی شده در بخش ۵ پرداخته می‌شود. سپس با آلوده کردن یک داده تولید شده از مدل پواسون توسط داده‌های بیش‌پراکنش، کارایی این مدل با سایر مدل‌ها مقایسه می‌شود. در ادامه نیز از این مدل برای تحلیل داده‌های واقعی استفاده می‌شود.

الگوریتم ۱: شبیه‌سازی و تولید داده از مدل پواسون- بیرنبام ساندرز با پارامترهای معلوم μ و σ به کمک قضیه ۱

گام ۱: تولید یک نمونه تصادفی nتایی از توزیع $\lambda \sim BS(\sigma^{\frac{1}{2}}, \frac{2}{\sigma+2})$ و نشان دادن آن‌ها با λ_i برای $i = 1, \ldots, n$.

گام ۲: محاسبه مقادیر $\mu\lambda_i$ برای مقدار معلوم μ و هر مقدار شبیه‌سازی شده λ_i.

گام ۳: تولید مقداری تصادفی از توزیع پواسون با پارامتر $\mu\lambda_i$ برای $i = 1, \ldots, n$.

لازم به ذکر است برای انجام گام ۱ می‌توان از روش‌های متعددی استفاده کرد. به عنوان نمونه می‌توان به بسته‌های bs و $GIGrvg$ در نرم‌افزار R اشاره نمود.

برای بررسی کارایی مدل پیشنهادی ابتدا نمونه‌های تصادفی ۵۰، ۱۰۰ و ۳۰۰ تایی از مدل پواسون با $\lambda = ۱$ شبیه‌سازی شده است. سپس ۳۰ درصد از داده‌ها به تصادف حذف شده و با نمونه‌ای تصادفی از توزیع پواسون با $\lambda = ۲۰$، که واریانس بزرگتری نسبت به داده‌های موجود دارد، تعویض شده‌اند. معیار ارزیابی عملکرد مدل پیشنهادی درصد قدرت تشخیص صحیح مدل در نظر گرفته شده است. هم‌چنین برای بررسی عملکرد مدل، مقادیر واقعی با مقادیر پیش‌بینی شده توسط مدل برتر بر اساس ملاک MSE مقایسه شده‌اند. با استفاده از نرم افزار R و بسته $gamlss$ مدل‌های مورد نظر به داده‌ها برازش داده شده و نتایج در جدول ۱ ارائه شده‌اند. همان‌طور که ملاحظه می‌شود با افزایش حجم نمونه کارایی مدل پیشنهادی

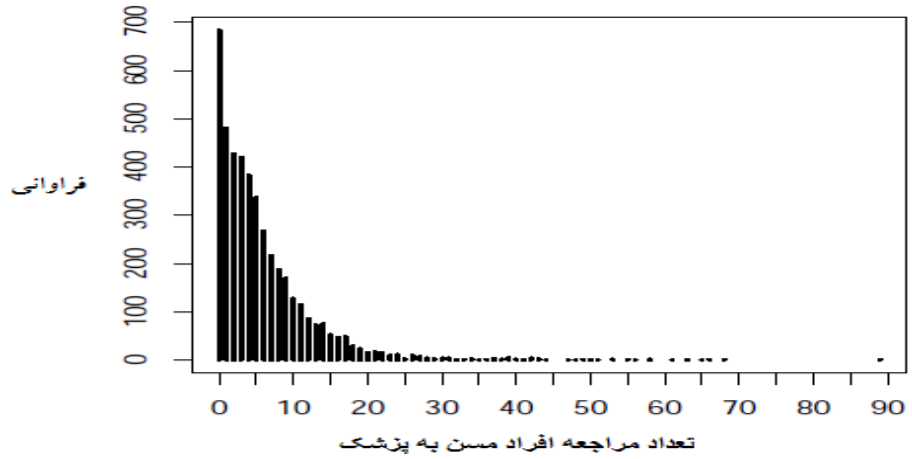
جدول ۱. مقادیر درصد قدرت تشخیص مدل و ملاک MSE با حجم نمونه‌های مختلف.

مدل	PO	PO-Gamma	PO-GIG	PO-BS	حجم نمونه
درصد تشخیص صحیح مدل	۰/۰۳	۰/۷۴۱	۰/۶۰۳	۰/۵۹۷	۵۰
	۰/۱۱	۰/۸۸۳	۰/۸۳۹	۰/۸۳۹	۱۰۰
	۰/۰۹	۰/۹۱۹	۰/۹۲۷	۰/۹۳۱	۳۰۰
میانگین MSE مدل	۰/۸۹	۰/۱۳۹	۰/۱۴۳	۰/۱۵۹	۵۰
	۰/۴۷	۰/۱۲۵	۰/۱۲۳	۰/۱۲۳	۱۰۰
	۰/۳۹	۰/۰۸۷	۰/۰۸۴	۰/۰۸۲	۳۰۰

از مدل‌های مرسوم بهتر می‌شود.

مثال ۱: (تقاضا برای مراقبت‌های پزشکی توسط افراد مسن) مجموعه داده‌های این مثال شامل اطلاعات مربوط به ۴۴۰۶ فرد ۶۶ ساله و بیشتر است که در یک برنامه بیمه عمومی پوشش داده شده و توسط دب و تریودی (۱۹۹۷) مورد تحلیل قرار گرفته‌اند و توسط زیلیس و کلیبر (۲۰۰۸) با عنوان داده‌های $DebTrivedi.rdi$ در پایگاه داده نرم‌افزار R بارگذاری شده‌اند.

برای مدل‌بندی میزان تقاضای مراقبت‌های پزشکی، تعداد مراجعه به پزشک به عنوان متغیر وابسته و تعداد اقامت در بیمارستان، وضعیت سلامت (ضعیف، متوسط و عالی)، تعداد بیماری‌های مزمن، جنسیت (زن و مرد)، تعداد سال تحصیلات و بیمه خصوصی (بلی و خیر) به عنوان متغیرهای توضیحی در نظر گرفته شده‌اند. نمودار فراوانی مراجعه بیماران به پزشک در شکل ۲ نشان داده شده است.



شکل ۲. فراوانی تعداد مراجعه به پزشک.

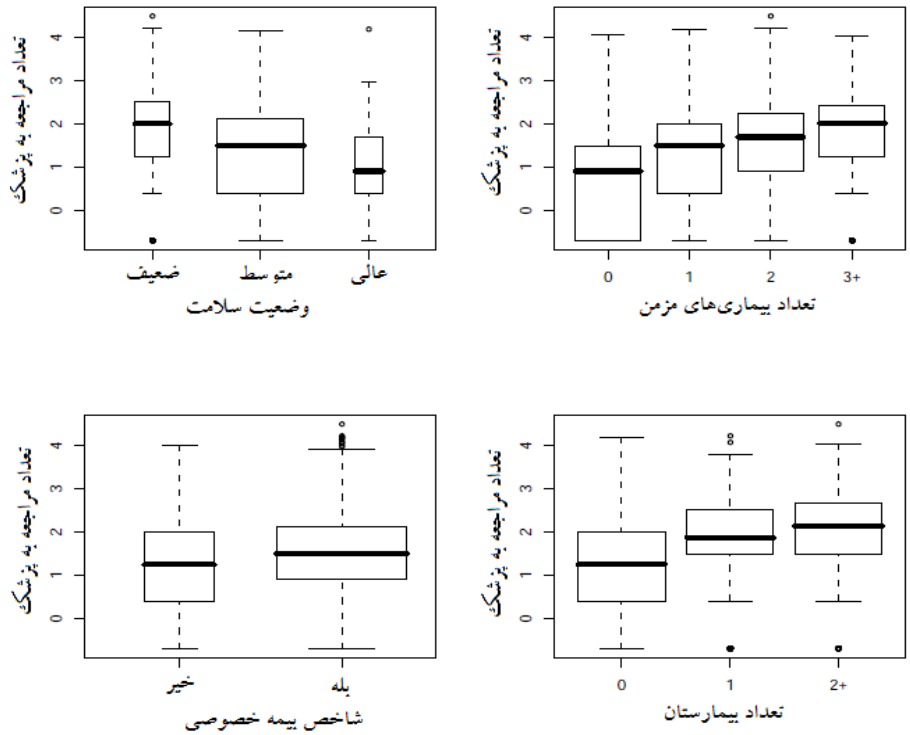
مقادیر میانگین و واریانس متغیر پاسخ به ترتیب ۵/۷۷۴ و ۴۵/۶۸۷ مشاهده شده‌اند که بیانگر بیش‌پراکنش داده‌ها است. بنابراین برای مدل‌بندی تعداد مراجعه افراد مسن به پزشک می‌توان از مدل‌های پواسون آمیخته استفاده کرد. همان‌طور که از شکل ۳ مشخص است متغیر پاسخ در مقابل متغیرهای توضیحی رفتاری صعودی یا نزولی از خود نشان داده است.

برای انتخاب بهترین مدل قابل برازش به داده‌ها، مدل‌های رگرسیون پواسون، پواسون- گاما، پواسون- گاوسی وارون و پواسون- بیرنبام ساندرز به داده‌ها برازش داده شده و با ملاک MSE و معیار اطلاع AIC مورد مقایسه قرار گرفته‌اند.

جدول ۲. مقادیر معیارهای MSE و AIC برای مدل‌های برازش شده

مدل	تعداد پارامترها	MSE	AIC
پواسون	۱	۲/۹۸	۳۵۹۵۹/۲۳
دو جمله‌ای منفی	۲	۲/۰۲	۲۴۳۵۹/۱۱
پواسون-گاوسی وارون	۲	۲/۱۶	۲۴۴۲۶/۹۳
پواسون- بیرنبام ساندرز	۲	۱/۹۶	۲۴۱۹۶/۷۲

با توجه به مقادیر MSE و معیار اطلاع اکائیکه در جدول ۲، مدل پواسون- بیرنبام ساندرز برازش بهتری را نسبت به سایر مدل‌ها از خود نشان داده است. در جدول ۳ ضرایب برآورد شده از مدل پواسون-



شکل ۳. رفتار متغیر وابسته در مقابل متغیرهای توضیحی

بیرنبام ساندرز ارائه شده‌اند.

با توجه به مقادیر t در جدول ۳ تأثیر متغیرهای توضیحی بر متغیر پاسخ کاملا مشخص است. طبق رابطه (۲) و جدول ۳ و با توجه به این‌که $E(\lambda) = ۱$ فرض شده است، مدل برازش شده برای μ توسط رگرسیون پواسون-بیرنبام ساندرز به‌صورت

$$E(Y_i|x_i) = \mu_i = \exp(۰/۸۶۷ + ۰/۲۰۵\mathbf{x}_۱ + ۰/۲۶۱\mathbf{x}_۲ - ۰/۳۵۴\mathbf{x}_۳ + ۰/۱۹۵\mathbf{x}_۴ - ۰/۱۴۶\mathbf{x}_۵ + ۰/۰۲۶\mathbf{x}_۶ + ۰/۲۳۵\mathbf{x}_۷)$$

نوشته می‌شود، که در آن $x_۱$ تعداد روز اقامت در بیمارستان، $x_۲$ وضعیت سلامت ضعیف، $x_۳$ وضعیت

جدول ۳. ضرایب برآورد شده از برازش مدل رگرسیون- پواسون بیرنبام ساندرز

ضرایب	برآورد	انحراف معیار	t-مقدار
β_0	۰/۸۶۷	۰/۰۵۲	۱۸/۷۲۵
β_1	۰/۲۰۵	۰/۰۱۹	۱۰/۹۵۳
β_2	۰/۲۶۱	۰/۰۴۴	۶/۷۸۳
β_3	-۰/۳۵۴	۰/۵۸۹	-۵/۷۶۲
β_4	۰/۱۹۵	۰/۰۱۲	۱۵/۲۳۱
β_5	-۰۱۴۶	۰/۰۳۱	-۳/۸۵۶
β_6	۰/۰۲۶	۰/۰۰۴	۶/۰۵۴
β_7	۰/۲۳۵	۰/۰۳۹	۶/۵۴۲

سلامت عالی، x_4 تعداد بیماری‌های مزمن، x_5 جنسیت مرد، x_6 تعداد سال تحصیلات و x_7 شاخص بیمه خصوصی بلی است. هم‌چنین $\hat{\sigma} = ۲/۴۲۱۵$.

بحث و نتیجه‌گیری

در این مقاله مدل رگرسیون پواسون-بیرنبام ساندرز معرفی شد که می‌تواند برای لحاظ مسئله بیش‌پراکنش در داده‌های شمارشی مورد استفاده قرارگیرد. از آن‌جا که توزیع بیرنبام ساندرز آمیخته‌ای از دو توزیع گاوسی وارون تعمیم‌یافته با مقادیر ثابت $\nu = \pm\frac{1}{2}$ است، بنابراین یک پارامتر کمتر نسبت به توزیع گاوسی وارون تعمیم‌یافته دارد. این مدل دو پارامتری کارایی خود را با یک پارامتر کمتر نسبت به مدل‌های قبلی بر اساس معیارهای MSE و AIC نشان داد. لذا استفاده از این مدل برای برازش و پیش‌بینی داده‌های شمارشی بیش‌پراکنش توصیه می‌شود.

تقدیر و تشکر

نویسندگان از نظرات ارزشمند داوران محترم در بهبود کیفیت مقاله و هم‌چنین از سردبیر، هیئت تحریریه و ویراستار مجله به خاطر مطالعه دقیق و ویرایش ادبی مقاله تقدیر و تشکر به عمل می‌آورند.

مراجع

Achcar, J. A. (1993), Inferences for the Birnbaum–Saunders Fatigue Life Model Using Bayesian Methods, *Computational Statistics and Data Analysis*, **15**, 367–380.

Birnbaum, Z. W. and Saunders, S. C. (1969), A New Family of Life Distribution, *Journal of Applied Probability*, **6**, 319-327.

Deb, P. and Trivedi, P. K. (1997), Demand for Medical Care by the Elderly: A Finite Mixture Approach, *Journal of Applied Econometrics*, **12**, 313-336.

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

Díaz-García, J. A. and Leiva, V. (2005), A New Family of Life Distributions Based on Elliptically Contoured Distributions, *Journal of Statistical Planning and Inference*, **128**, 445–457.

From, S. G. and Li, L. X. (2006), Estimation of the Parameters of the Birnbaum Saunders Distribution, *Communications in Statistics, Theory and Methods*, **35**, 2157-2169.

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.

Gradshteyn, I. S. and Ryzhik, I. M. (2000), *Table of Integrals, Series and Products*, 6th Edt., Academic Press, San Diego.

Hashimotoa, E. M., Ortegaa, E. M. M., Cordeirob, G. M. and Canchoc, V. G. (2013), The Poisson Birnbaum–Saunders Model with Long-term Survivors, *A Journal of Theoretical and Applied Statistics*, **48**, 1394–1413.

Hinde, J. (1982), *Compound Poisson Regression Models*, In: Gilchrist, R. (Ed.), GLIM 82, Proceedings of the International Conference on Generalised Linear Models, Springer, New York, 109-121.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992), *Univariate Distributions*, 2nd Edt., John Wiley, New York.

Karlis, D. (2005), EM Algorithm for Mixed Poisson and Other Discrete Distributions, *Astin Bulletin*, **35**, 3-24.

Karlis, D. and Xekalaki, E. (2005), Mixed Poisson Distributions, *International Statistical Review*, **73**, 35-58.

Kundu, D. N., Balakrishnan, N. and Jamalizadeh, A. (2010), Bivariate Birnbaum Saunders Distribution and Associated Inference, *Journal of Multivariate Analysis*, **101**, 113-125.

Lemonte, A. J., Cribari-Neto, F. and Vasconcellos, K. L. P. (2007), Improved Statistical Inference for the Two-Parameter Birnbaum Saunders Distribution, *Computatinal Statistics and Data Analysis*, **51**, 4656-4681.

Neal, P. and Roberts, G. (2008), Optimal Scaling for Random Walk Metropolis on Spherically Constrained Target Densities, *Methodology and Computing in Applied Probability*, **10**, 277–297.

Miaou, S. P. (1994), The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions, *Accident Analysis and Prevention*, **26**, 471-482.

Patil, G. P. (1970), *Random Counts in Models and Structures*, Pennsylvania State University Press, University Park, London.

Rigby, R. A., Stasinopoulos, D. M. and Akantziliotou, C. (2008), A Framework for Modelling Overdispersed Count Data, Including the Poisson-shifted Generalized Inverse Gaussian Distribution, *Statistics and Data Analysis*, **53**, 381-393.

Saez-Castillo, A. J. and Cond-Sanchez, A. (2013), A Hyper-Poisson Regression Model for Overdispersed and Underdispersed Coun Data, *Compulational Statistics and Data Analisis*, **61**, 146-157.

Sapatinas, T. (1995), Identifiability of Mixtures of Power-series Distributions and Related Characterizations, *Annals of the Institute of Statistical Mathematics*, **47**, 447-459.

Stein, G. Z. and Juritz, J. M. (2007), Linear Models with an Inverse Gaussian Poisson Error Distribution, *Communications in Statistics, Theory and Methods*, **17**, 557-571.

Taylor, H. M. and Karlin, S. (1994), *An Introduction to Stochastic Modelling*, Revised Edition, Academic Press, San Diego and New York.

Wang, M., Sun, X. and Park, C. (2016), Bayesian Analysis of Birnbaum–Saunders Distribution via the Generalized Ratio-of-Uniforms Method, *Computational Statistics*, **31**, 207-225 .

Xie, F. C. and Wei, B. C. (2008), Influence Analysis for Poisson Inverse Gaussian Regression Models Based on the EM Algorithm, *Metrika*, **67**, 49-62.

Xu, A. and Tang, Y. (2011), Bayesian Analysis of Birnbaum–Saunders Distribution with Partial Information, *Computational Statistics and Data Analysis*, **55**, 2324-2333.

Zeileis, A. and Kleiber, C. (2008), **AER**: *Applied Econometrics with R*, R Package Version 0.9-0, URL http://CRAN.R-project.org/package=AER.