

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

رضا پورموسی و نرجس گیلانی

بخش آمار، دانشکده ریاضی و کامپیوتر، دانشگاه شهید باهنر کرمان

**چکیده:** در این مقاله ابتدا به معرفی مدل‌های رگرسیون پواسون آمیخته پرداخته و در ادامه به معرفی یک مدل جدید به نام رگرسیون پواسون-بیرنهام ساندرز با هدف لحاظ کردن مسئله بیش‌پراکنش در مدل‌بندی داده‌های شمارشی پرداخته می‌شود. از آنجا که توزیع بیرنهام 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:** توزیع بیرنهام 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 حالی که در بعضی از توزیع‌ها مانند 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<sup>۱</sup> و کم‌پراکنش<sup>۲</sup> 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 و آرون، پواسون-گاوسی و آرون تعمیم‌یافته مثال‌هایی از این نتایج هستند (گاردنر و همکاران، ۱۹۹۵؛ استین و جوریتز، ۲۰۰۷؛ رایگی و همکاران، ۲۰۰۸).

مسئله بیش‌پراکنش زمانی رخ می‌دهد که پارامتر توزیع پواسون به تنهایی قادر به توصیف رخداد‌های شمارشی نیست.

<sup>۱</sup>Overdispersion

<sup>۲</sup>Underdispersion

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<sup>۳</sup> برای افزایش کارایی مدل در داده‌های بیش‌پراکنش است. برای این منظور در بخش‌های ۲ و ۳ مدل رگرسیون پواسون، توزیع پواسون آمیخته و خواص آن‌ها معرفی شده و مدل‌های رگرسیون پواسون-گاما<sup>۴</sup>، پواسون-گوسی وارون<sup>۵</sup> و پواسون-گوسی وارون تعمیم‌یافته<sup>۶</sup> که از این خانواده نتیجه شده‌اند به همراه ویژگی‌های آن‌ها تعریف شده‌اند. در بخش ۴ مدل جدیدی از رگرسیون پواسون آمیخته به نام پواسون-بیرنباوم ساندرز<sup>۷</sup> 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 ۰.۱** (مدل رگرسیون پواسون) فرض کنید  $Y_i$  دارای توزیع پواسون با میانگین  $\mu_i$  باشد، آن‌گاه:

$$f_{Y_i}(y_i | x_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \quad \mu_i > 0, \quad y_i = 0, 1, \dots$$

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:

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

<sup>۳</sup>Mixed Poisson regression models

<sup>۴</sup>Poisson-Gamma

<sup>۵</sup>Poisson- inverse Gaussian

<sup>۶</sup>Poisson-generalized inverse Gaussian

<sup>۷</sup>Poisson- Birnbaum Saunders

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

$$E(Y_i | \mathbf{x}_i) = \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}), \quad (2)$$

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

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

هستند. به منظور مدل‌بندی داده‌های شمارشی معمولاً فرض

$$\text{Var}(Y) = \phi E(Y) = \phi \mu,$$

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

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

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

**گزاره ۰۱.** اگر متغیر تصادفی  $Y$  دارای توزیع پواسون آمیخته باشد، آن‌گاه میانگین و واریانس آن عبارتند از

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

$$\text{Var}(Y) = \mu E(\lambda) + \mu^2 \text{Var}(\lambda).$$

طبق رابطه (۲) چون پارامتر  $\mu$  نقش مدل رگرسیون پواسون را به عهده دارد، لذا با استفاده از گزاره ۰۱، در توزیع‌های

پواسون آمیخته، پارامترهای توزیع آمیزنده را طوری در نظر می‌گیرند که  $E(\lambda) = 1$  شود. در این صورت

$$E(Y) = \mu, \quad (3)$$

$$Var(Y) = \mu + \mu^2 Var(\lambda). \quad (4)$$

با توجه به رابطه (۴) واریانس توزیع پواسون آمیخته همواره بزرگتر از واریانس توزیع پواسون است. معمولاً 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

$$f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \quad y > 0,$$

$$f_Y(y) = \sqrt{\frac{\beta}{\pi y^3}} \exp\left\{-\frac{\beta(y-\alpha)^2}{2\alpha^2 y}\right\}, \quad y > 0,$$

$$f_Y(y) = \frac{\left(\frac{\psi}{\chi}\right)^{\frac{\nu}{2}}}{\sqrt{K_\nu(\sqrt{\psi\chi})}} y^{\nu-1} \exp\left(-\frac{1}{2}(\chi y^{-1} + \psi y)\right), \quad y > 0.$$

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

با انتخاب توزیع‌های آمیزنده متفاوت، می‌توان توزیع‌های پواسون آمیخته جدیدی را تعریف نمود، اما 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  $\lambda$  دارای توزیع پواسون با میانگین  $\mu\lambda$  باشد، که در آن  $\mu = \exp(\mathbf{x}^T \boldsymbol{\beta})$  و  $\lambda \sim \Gamma\left(\frac{1}{\sigma}, \frac{1}{\sigma}\right)$ . در این صورت تابع

$$P(Y = y) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(y + 1) \Gamma(\frac{1}{\sigma})} \left(\frac{\sigma\mu}{1 + \sigma\mu}\right)^y \left(\frac{1}{1 + \sigma\mu}\right)^{\frac{1}{\sigma}}, \quad y = 0, 1, \dots$$

با میانگین و واریانس

$$E(Y) = \mu, \quad Var(Y) = \mu(1 + \sigma\mu)$$

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

پواسون و رگرسیون پواسون-گاما پرداخته‌اند.

**تعریف ۰۴.** (مدل رگرسیون پواسون-گاوسی وارون،  $PO-IG$ ) فرض کنید توزیع  $Y$  به شرط متغیر تصادفی مشاهده نشده  $\lambda$  پواسون با میانگین  $\mu\lambda$  باشد، که در آن  $\mu = \exp(\mathbf{x}^T\boldsymbol{\beta})$  و  $\lambda \sim IG(1, \frac{1}{\sigma})$ . در این صورت تابع جرم احتمال حاشیه‌ای  $Y$  به صورت

$$P(Y = y) = \left(\frac{\nu\theta}{\pi}\right)^{\frac{1}{2}} \frac{\mu^y}{y!} e^{-\frac{1}{\sigma}} \frac{K_{y-\frac{1}{2}}(\theta)}{(\theta\sigma)^y}, \quad y = 0, 1, \dots$$

است، که در آن  $\theta^2 = \frac{1}{\sigma^2} + \frac{\nu\mu}{\sigma}$  و واریانس آن عبارتند از

$$E(Y) = \mu, \quad Var(Y) = \mu(1 + \sigma\mu).$$

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

**تعریف ۰۵.** (مدل رگرسیون پواسون-گاوسی وارون تعمیم‌یافته،  $PO-GIG$ ) فرض کنید توزیع  $Y$  به شرط متغیر تصادفی مشاهده نشده  $\lambda$  پواسون با میانگین  $\mu\lambda$  باشد، که در آن  $\mu = \exp(\mathbf{x}^T\boldsymbol{\beta})$  و  $\lambda \sim GIG(\nu, \frac{1}{c\sigma}, \frac{c}{\sigma})$ . در این صورت تابع جرم احتمال حاشیه‌ای  $Y$  به صورت

$$P(Y = y) = \left(\frac{\mu}{c}\right)^y \frac{1}{(\theta\sigma)^{y+\nu}y!} \frac{K_{\nu+y}(\theta)}{K_{\nu}(\frac{1}{\sigma})}, \quad y = 0, 1, \dots$$

است، که در آن

$$\theta^2 = \frac{1}{\sigma^2} + \frac{\nu\mu}{c\sigma}, \quad c = R_{\nu}\left(\frac{1}{\sigma}\right) = \frac{K_{\nu+1}\left(\frac{1}{\sigma}\right)}{K_{\nu}\left(\frac{1}{\sigma}\right)}$$

و میانگین و واریانس آن عبارتند از

$$E(Y) = \mu, \quad Var(Y) = \mu\left(1 + \mu\left(\frac{\nu\sigma(\nu+1)}{c} + \frac{1}{c^2} - 1\right)\right).$$

رایگی و همکاران (۲۰۰۸) به معرفی ویژگی‌های توزیع پواسون-گاوسی وارون تعمیم‌یافته و چارچوبی برای

مدلبندی داده‌های شمارشی پرداخته‌اند.

## ۴ معرفی مدل رگرسیون پواسون-بیرنباام ساندرز

در این بخش مدلی از خانواده رگرسیون پواسون آمیخته به نام پواسون-بیرنباام ساندرز معرفی می‌شود، که در آن  $\lambda$  دارای توزیع آمیزنده بیرنباام ساندرز فرض شده است. این ایده از ارتباط توزیع گاوسی وارون تعمیم‌یافته و توزیع بیرنباام 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 بیرنجام ساندرز ترکیب خطی از دو توزیع گاوسی وارون تعمیم‌یافته است، به نظر می‌رسد که انتخاب توزیع بیرنجام 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-بیرنجام ساندرز به منظور مدل‌بندی میزان بهبودی بقا و با این فرض که تعداد پیشامدها از توزیع پواسون و زمان پیشامدها از توزیع بیرنجام ساندرز پیروی می‌کنند، معرفی کرده‌اند. ایده ارائه شده توسط آن‌ها کاملاً متفاوت با ایده پیشنهادی در این مقاله است.

## ۱.۴ توزیع بیرنجام 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 قرار گرفته و پس از یک دوره کار زیاد ترک خورده و کم‌کم می‌شکنند. این اتفاق معمولاً بدون اطلاع قبلی و غیر قابل رویت رخ می‌دهد. توزیع بیرنجام ساندرز که برای نخستین بار توسط **بیرنجام 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 **دیزگارسیا و لیوا (۲۰۰۵)**، **فرام و لی (۲۰۰۶)** و **لیمون و همکاران (۲۰۰۷)** به بسط و تعمیم این توزیع پرداخته‌اند. همچنین **کاندو و همکاران (۲۰۱۰)** توزیع بیرنجام ساندرز دو بعدی که دارای توزیع‌های حاشیه‌ای بیرنجام ساندرز تک بعدی است را معرفی کرده‌اند. این توزیع از دیدگاه بیزی ابتدا توسط **آچکار (۱۹۹۳)** مورد بررسی قرار گرفت، 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 بیرنهام ساندرز  $BS(\alpha, \beta)$  است، اگر تابع توزیع تجمعی آن به صورت

$$F_{BS}(w; \alpha, \beta) = \Phi\left[\frac{1}{\alpha}\left\{\left(\frac{w}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{w}\right)^{\frac{1}{2}}\right\}\right], \quad w, \alpha, \beta > 0,$$

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

$$f_{BS}(w; \alpha, \beta) = \frac{1}{\sqrt{2\pi\alpha\beta}} \left[ \left(\frac{\beta}{w}\right)^{\frac{1}{2}} + \left(\frac{w}{\beta}\right)^{\frac{1}{2}} \right] \exp\left[-\frac{1}{2\alpha^2} \left(\frac{w}{\beta} + \frac{\beta}{w} - 2\right)\right]. \quad (5)$$

**گزاره ۲.** امید ریاضی و واریانس توزیع بیرنهام ساندرز عبارتند از

$$E(W) = \beta\left(1 + \frac{1}{\alpha^2}\right), \quad \text{Var}(W) = (\alpha\beta)^2\left(1 + \frac{5}{\alpha^2}\right).$$

یکی از خواص مورد توجه در توزیع بیرنهام ساندرز این است که می‌توان آن را به صورت آمیخته‌ای از دو توزیع گاوسی وارون تعمیم‌یافته (اتینه هالفن، ۱۹۴۶) 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 به صورت آمیخته‌ای از دو متغیر تصادفی گاوسی وارون تعمیم‌یافته مستقل از هم به صورت

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

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

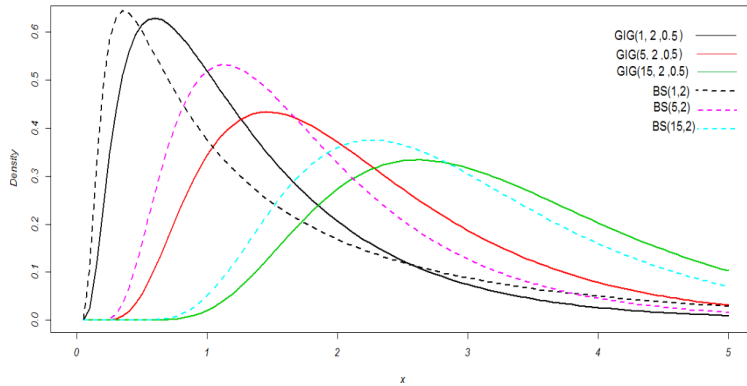
$$f_{BS}(w; \alpha, \beta) = \frac{1}{2} f_{GIG}\left(w; \frac{1}{2}, \frac{\beta}{\alpha^2}, \frac{1}{\beta\alpha^2}\right) + \frac{1}{2} f_{GIG}\left(w; -\frac{1}{2}, \frac{\beta}{\alpha^2}, \frac{1}{\beta\alpha^2}\right).$$

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

## ۲.۴ مدل رگرسیون پواسون-بیرنهام ساندرز

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





شکل ۱: نمودار تابع چگالی توزیع‌های گاوسی وارون تعمیم‌یافته و بیرنجام استاندارد برای چند ترکیب مختلف از پارامترها

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

$$f_Y(y) = \left(\frac{\mu\beta}{\alpha^\nu \delta}\right)^y \frac{e^{-\frac{1}{\alpha^\nu}}}{\sqrt{\nu\pi y!} \alpha} [\alpha^{-1} \delta^{-\frac{1}{\nu}} K_{y+\frac{1}{\nu}}(\delta) + \alpha \delta^{\frac{1}{\nu}} K_{y-\frac{1}{\nu}}(\delta)], \quad y = 0, 1, \dots \quad (6)$$

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

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

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

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

$$\begin{aligned} &\int_0^\infty \lambda^{y-\frac{1}{\nu}} e^{-\frac{1}{\nu} \left[ \left(\frac{1}{\beta\alpha^\nu} + \nu\mu\right)\lambda + \frac{\beta}{\alpha^\nu} \lambda^{-1} \right]} \\ &d\lambda = \nu \left( \frac{\alpha^\nu \delta}{\beta} \right)^{-y-\frac{1}{\nu}} K_{y+\frac{1}{\nu}}(\delta) \end{aligned}$$

به دست می‌آیند. هم‌چنین

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

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

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

که از آن رابطه (۶) حاصل می‌شود.

□

در قضیه ۱ برای اعمال فرض  $E(\lambda) = 1$ ، طبق رابطه (۳) بایستی پارامترهای  $\alpha$  و  $\beta$  در توزیع بیرنهام ساندرز را به صورت  $\alpha = \sigma^{\frac{\gamma}{\beta}}$  و  $\beta = \frac{\gamma}{\sigma + \gamma}$  در نظر گرفت.

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

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

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

$$Var(Y) = \mu + \mu^{\gamma} \left( \frac{\sigma(\gamma + 5\sigma)}{(\sigma + \gamma)^{\gamma}} \right).$$

□

بنابراین این مدل ویژگی (۴) را دارا است.

## ۵ برآورد ماکسیمم درست‌نمایی مدل پواسون-بیرنهام ساندرز

به‌دست آوردن برآورد ماکسیمم درست‌نمایی پارامترهای مدل رگرسیون پواسون-بیرنهام ساندرز مستلزم حل معادلات غیر خطی با روش‌های عددی تکراری است. الگوریتم امید-ماکسیمم  $EM^{\wedge}$  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.

<sup>^</sup>EM algorithm

الگوریتم  $EM$  معمولاً برای مواردی که تابع درست‌نمایی شکل پیچیده‌ای دارد و برآورد بیشینه درست‌نمایی پارامترها با روش مستقیم کار غیرممکن است، مورد استفاده قرار می‌گیرد. در مدل رگرسیون پواسون بیرنهام ساندرز، فرض کنید  $\theta = (\beta^T, \sigma)^T$  بردار پارامترهای نامعلوم،  $Y_{obs}$  نشان دهنده داده‌های مشاهده شده  $\{Y_i; i = 1, \dots, n\}$  و  $\{\lambda_i; i = 1, \dots, n\}$  دنباله‌ای از وزن‌ها باشند که به عنوان داده‌های گمشده  $Y_{mis}$  رفتار می‌کنند. به علاوه  $Y_c = (Y_{obs}, Y_{mis})$  اشاره به مجموعه داده‌های کامل دارد.

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

۱. گام  $E$ ، مرحله گرفتن امید ریاضی: در این گام امید ریاضی لگاریتم تابع درست‌نمایی داده‌های کامل به شرط داده‌های مشاهده شده و مقدار پارامتر جاری محاسبه می‌شود. به عبارتی

$$E(\log L(Y_c|\theta)|y_{obs}, \theta^{(k)}) = Q(\theta|\theta^{(k)})$$

محاسبه می‌شود.

۲. گام  $M$ ، مرحله ماکسیم‌سازی: در این گام،  $\theta^{(k+1)}$  به گونه‌ای انتخاب می‌شود که تابع  $Q(\theta|\theta^{(k)})$  برای هر  $\theta$  ماکسیم شود، به عبارت دیگر

$$\theta^{(k+1)} = \underset{\theta}{\text{Arg max}}(Q(\theta|\theta^{(k)})).$$

گام  $E$  و گام  $M$  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<sup>۹</sup> استفاده می‌شود.

<sup>۹</sup>Power series family

**تعریف ۰.۷.** یک توزیع گسسته متعلق به خانواده توزیع‌های سری توانی نامیده می‌شود، اگر تابع جرم احتمال آن را بتوان

به صورت

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

نشان داد، که در آن  $a_y > 0$  و  $A(\lambda)$  تابعی از  $\lambda$  است و به  $y$  بستگی ندارد. بسیاری از توزیع‌های گسسته شناخته شده مانند توزیع پواسون و دو جمله‌ای متعلق به این خانواده هستند. اگر پارامتر  $\lambda$  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 ۰.۱.** (سایپاتیناس، ۱۹۹۵) اگر  $Y|\lambda$  از توزیع‌های گسسته توانی پیروی کند و  $\lambda$  دارای تابع احتمال حاشیه‌ای  $g(\lambda)$

باشد، آن‌گاه

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

**قضیه ۰.۳.** فرض کنید  $Y$  به شرط  $\lambda$  دارای توزیع پواسون با میانگین  $\mu\lambda$  و  $\lambda \sim BS(\alpha, \beta)$  باشد، آن‌گاه

الف: تابع چگالی  $\lambda$  به شرط مشاهده  $y$  به صورت

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

است، که در آن

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

ب: گشتاور  $r$ ام متغیر تصادفی  $\lambda$  به شرط مشاهده  $y$  برابر است با:

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

برهان. الف: با توجه به این‌که

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

که در آن  $k$  تابعی از  $y$  و مستقل از  $\lambda$  است. بنابراین

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

با توجه به خواص تابع بسط تعمیم یافته نوع سوم می توان نوشت:

$$c = k' = \frac{1}{\sqrt{\alpha}\beta} \left(\frac{\delta\alpha^{\alpha}}{\beta}\right)^y \left[(\delta\alpha^{\alpha})^{-\frac{1}{\alpha}} K_{y+\frac{1}{\alpha}}(\delta) + (\delta\alpha^{\alpha})^{\frac{1}{\alpha}} K_{y-\frac{1}{\alpha}}(\delta)\right]^{-1}$$

بنابراین

$$f(\lambda|y) = c[\beta^{\frac{1}{\alpha}} \lambda^{y-\frac{1}{\alpha}} e^{-\frac{1}{\alpha}\left(\left(\frac{1}{\beta\alpha^{\alpha}} + \alpha\mu\right)\lambda + \frac{\beta}{\lambda\alpha^{\alpha}}\right)} + \beta^{\frac{\alpha}{\alpha}} \lambda^{y-\frac{\alpha}{\alpha}} e^{-\frac{1}{\alpha}\left(\left(\frac{1}{\beta\alpha^{\alpha}} + \alpha\mu\right)\lambda + \frac{\beta}{\lambda\alpha^{\alpha}}\right)}].$$

برای اثبات (ب) می توان از قضیه ۳ یا ویژگی خانواده سری های توانی استفاده کرد. با توجه به این که توزیع پواسون

متعلق به خانواده توزیع های سری توانی با  $a_y = \frac{\mu^y}{y!}$ ,  $A(\lambda) = e^{\mu\lambda}$  است، با استفاده از لم ۱ اثبات به سادگی انجام

□

می شود.

رابطه (۷) برای برآورد پارامترهای مدل رگرسیون پواسون-بیرنهام ساندرز به کمک الگوریتم  $EM$  مورد استفاده

قرار می گیرد. تابع درست نمایی یک نمونه از توزیع پواسون-بیرنهام ساندرز با پارامترهای  $\theta = (\beta^T, \sigma)^T$  به صورت

$$\log L_Y(\theta) = \sum_{i=1}^n \log f_{y_i}(y_i; \theta).$$

است. تابع درست‌نمایی داده‌های کامل به صورت

$$\begin{aligned}\log L_{Y,\lambda}(\theta) &= \log \prod_{i=1}^n f_{Y_i|\lambda_i}(y_i; \theta) f_{\lambda_i}(\lambda_i, \theta) \\ &= \sum_{i=1}^n \log f_{Y_i|\lambda_i}(y_i; \theta) + \sum_{i=1}^n \log f_{\lambda_i}(\lambda_i; \theta) \\ &= L_{Y_n|\lambda_n}^{(1)}(\theta) + L_{\lambda_n}^{(2)}(\theta),\end{aligned}$$

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

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

این روابط در الگوریتم  $EM$  مورد استفاده قرار می‌گیرند.

## ۶ مثال‌های کاربردی

در این بخش ابتدا به شبیه‌سازی از مدل پواسون-بیرنهام 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-بیرنهام ساندرز با پارامترهای معلوم  $\mu$  و  $\sigma$  به کمک قضیه ۱

گام ۱: تولید یک نمونه تصادفی  $n$  تایی از توزیع  $BS(\sigma^{\frac{1}{\varphi}}, \frac{\varphi}{\sigma + \varphi})$  و نشان دادن آن‌ها با  $\lambda_i$  برای  $i =$

$1, \dots, n$ .

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

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

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  $\lambda = 1$  شبیه‌سازی شده است. سپس ۳۰ درصد از داده‌ها به تصادف حذف شده و با نمونه‌ای تصادفی از توزیع پواسون با  $\lambda = 20$ ، 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 نرم‌افزار *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

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

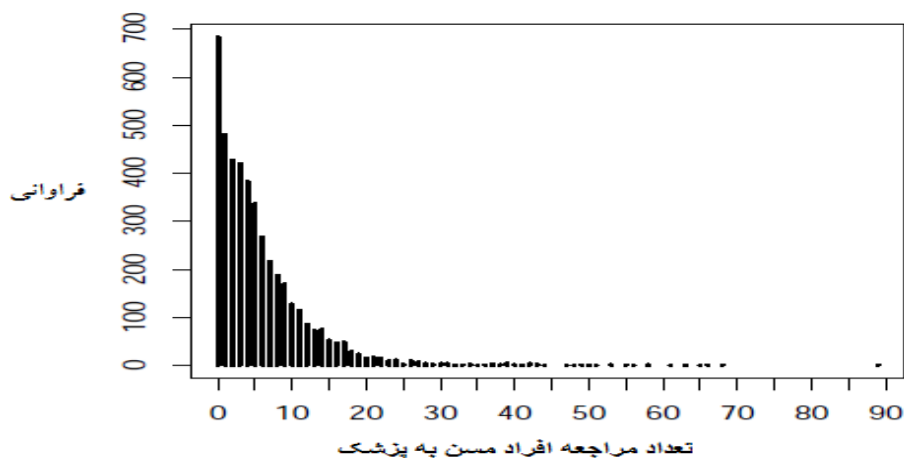
حجم نمونه	PO-BS	PO-GIG	PO-Gamma	PO	مدل
۵۰	۰/۵۹۷	۰/۶۰۳	۰/۷۴۱	۰/۰۳	
۱۰۰	۰/۸۳۹	۰/۸۳۹	۰/۸۸۳	۰/۱۱	درصد تشخیص صحیح مدل
۳۰۰	۰/۹۳۱	۰/۹۲۷	۰/۹۱۹	۰/۰۹	
۵۰	۰/۱۵۹	۰/۱۴۳	۰/۱۳۹	۰/۸۹	
۱۰۰	۰/۱۲۳	۰/۱۲۳	۰/۱۲۵	۰/۴۷	میانگین <i>MSE</i> مدل
۳۰۰	۰/۰۸۲	۰/۰۸۴	۰/۰۸۷	۰/۳۹	

حجم نمونه کارایی مدل پیشنهادی از مدل‌های مرسوم بهتر می‌شود.

**مثال ۰۱.** (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 کلیپر (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.



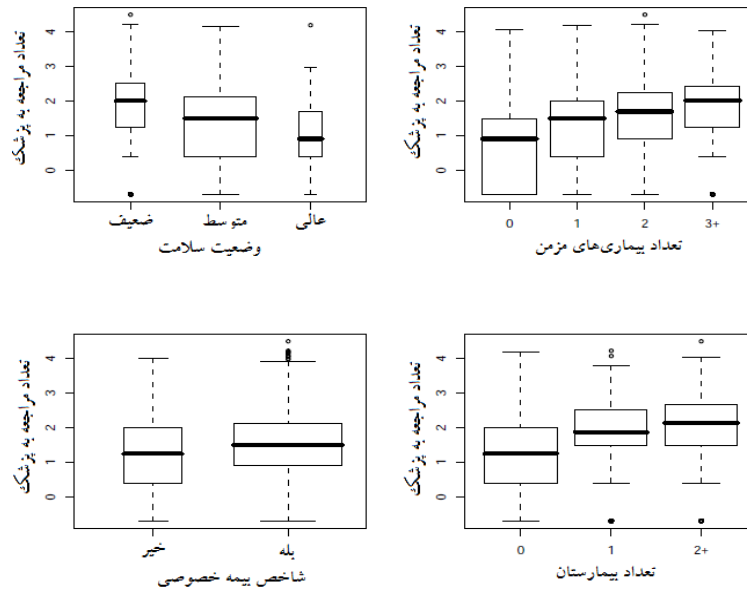
شکل ۲: فراوانی تعداد مراجعه به پزشک.

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 و وارون پواسون-بیرنهام ساندرز به داده‌ها برازش شده و با ملاک  $MSE$  و معیار اطلاع  $AIC$  مورد مقایسه قرار گرفته‌اند.

با توجه به مقادیر  $MSE$  و معیار اطلاع اکائیکه در جدول ۲، مدل پواسون-بیرنهام ساندرز برازش بهتری را نسبت به سایر مدل‌ها از خود نشان داده است. در جدول ۳ ضرایب برآورد شده از مدل پواسون-بیرنهام ساندرز ارائه شده‌اند. با توجه به مقادیر  $t$  در جدول ۳ تأثیر متغیرهای توضیحی بر متغیر پاسخ کاملاً مشخص است. طبق رابطه (۲) و جدول ۳ و با توجه به این‌که  $E(\lambda) = ۱$  فرض شده است، مدل برازش شده برای  $\mu$  توسط رگرسیون پواسون-بیرنهام





شکل ۳: رفتار متغیر وابسته در مقابل متغیرهای توضیحی

ساندرز به صورت

$$E(Y_i|x_i) = \mu_i = \exp(0/867 + 0/205x_1 + 0/261x_2 - 0/354x_3 + 0/195x_4 - 0/146x_5 + 0/26x_6 + 0/235x_7)$$

نوشته می‌شود، که در آن  $x_1$  تعداد روز اقامت در بیمارستان،  $x_2$  وضعیت سلامت ضعیف،  $x_3$  وضعیت 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  $\hat{\sigma} = 2/4215$ .

## بحث و نتیجه‌گیری

در این مقاله مدل رگرسیون پواسون-بیرنهام ساندرز معرفی شد که می‌تواند برای لحاظ مسئله بیش‌براکنش در داده‌های شمارشی مورد استفاده قرار گیرد. از آن‌جا که توزیع بیرنهام ساندرز آمیخته‌ای از دو توزیع گاوسی و آرون تعمیم‌یافته با مقادیر ثابت  $\nu = \pm \frac{1}{2}$  است، بنابراین یک پارامتر کمتر نسبت به توزیع گاوسی و آرون تعمیم‌یافته دارد. این مدل دو

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-گوسی وارون	۲	۲/۱۶	۲۴۴۲۶/۹۳
پواسون-بیرنهام استاندارد	۲	۱/۹۶	۲۴۱۹۶/۷۲

جدول ۳: ضرایب برآورد شده از برازش مدل رگرسیون-پواسون بیرنهام استاندارد

ضرایب	برآورد	انحراف معیار	$t$ -مقدار
$\beta_0$	۰/۸۶۷	۰/۰۵۲	۱۸/۷۲۵
$\beta_1$	۰/۲۰۵	۰/۰۱۹	۱۰/۹۵۳
$\beta_2$	۰/۲۶۱	۰/۰۴۴	۶/۷۸۳
$\beta_3$	-۰/۳۵۴	۰/۵۸۹	-۵/۷۶۲
$\beta_4$	۰/۱۹۵	۰/۰۱۲	۱۵/۲۳۱
$\beta_5$	-۰/۱۴۶	۰/۰۳۱	-۳/۸۵۶
$\beta_6$	۰/۰۲۶	۰/۰۰۴	۶/۰۵۴
$\beta_7$	۰/۲۳۵	۰/۰۳۹	۶/۵۴۲

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, *Comput Stat Data Anal*, **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, *Computational Statistics & Data Analysis*, **51**, 4656-4681.

Neal, P. and Roberts, G. (2008), Optimal Scaling for Random Walk Metropolis on Spherically Constrained Target Densities, *Methodol Comput Appl Probab*, **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, *Computational Statistics and Data Analysis*, **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, *Comput Stat*, **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, *Comput Stat Data Anal*, **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>.

# **Modeling Count Data Under the Influence Overdispersion by Poisson Birnbaum-Saunders Regression Model**

**Pourmousa, R. and Gilani, N.**

Department of Statistics, Shahid Bahonar University, Kerman, Iran.

**Abstract:** In this paper the mixed Poisson regression model is discussed and a Poisson Birnbaum-Saunders regression model is introduced consider the over-dispersion. The Birnbaum-Saunders distribution is the mixture of two the generalized inverse Gaussian distributions, therefore it can be considered as an extension of traditional models. Our proposed model has less dimensional parameter space than the Poisson- generalized inverse Gaussian regression model. We also show that the proposed model has a closed form for likelihood function and we obtain its moments. The EM algorithm is used to estimate the parameters and its efficiency is compared with conventional models by a simulation study. An analysis of a real data is provided for more illustration.

**Keywords:** Birnbaum-Saunders distribution, EM algorithm, Count data, Poisson regression models, Overdispersion, Mixed Poisson regression model.

**Mathematics Subject Classification (2010):** 62E15, 62J12