مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۲، شماره‌ی ۱، بهار و تابستان ۱۳۹۰، صص ۱۷ – ۳۲

مقایسه‌ی انواع مدل‌های رگرسیون در تحلیل داده‌های شمارشی

فاطمه حسن‌زاده* و ایرج کاظمی

دانشگاه اصفهان

چکیده: مدل متداول در برازش داده‌های شمارشی پواسون است که ویژگی مهم آن برابری میانگین و واریانس می‌باشد. اما در مباحث کاربردی در صورت عدم برقراری این شرط استفاده از مدل پواسون منجر به استنباط نادرستی از پارامترهای مدل خواهد شد. در این مقاله ضمن بررسی موضوع بیش‌پراکنش و آزمون‌های مربوط به آن، مدل‌های دوجمله‌ای منفی و پواسون تعمیم‌یافته به‌عنوان جایگزین‌هایی برای مدل پواسون معرفی‌می‌شوند. در ادامه، به مقایسه‌ی مدل‌های مختلف رگرسیونی برای داده‌های شمارشی با شرط ثابت بودن دو گشتاور مرتبه‌ی اول می‌پردازیم. همچنین، روش ماکسیمم درستنمایی را برای براورد پارامترهای مدل‌های رگرسیونی نام‌برده منظور می‌کنیم. در آخر، مدل‌های متفاوتی را به داده‌های اثر اجزای سندروم متابولیک در کودکان چاق شهر اصفهان بر فاکتور تعداد گلبول‌های سفید خون برازش داده و از بین آن‌ها بهترین را که برازنده‌ی داده‌هاست بر اساس معیارهای متداول انتخاب می‌کنیم.

واژگان‌کلیدی: روش ماکسیمم درستنمایی؛ مدل پواسون؛ دوجمله‌ای منفی؛ مسئله‌ی بیش‌پراکنش؛ پواسون تعمیم‌یافته؛ معیارهای اطلاع آکائیکی و بیزی.

۱- مقدمه

تحلیل داده‌های شمارشی در شاخه‌های مختلف علوم از جمله اقتصاد، علوم اجتماعی و پزشکی کاربرد دارد. مدل متداول برای برازش داده‌های شمارشی، مدل رگرسیون پواسون است. ویژگی مهم آن برابری میانگین و واریانس است. اما در بسیاری از تحلیل‌ها این شرط همواره برقرار نبوده و با مسئله‌ی بیش‌پراکنش و یا کم‌پراکنش مواجه می‌شویم.

---

* نویسنده‌ی عهده‌دار مکاتبات

دریافت: ۱۳۸۹/۵/۶، پذیرش: ۱۳۹۰/۱/۳۰.

هری و رانگ [۴] به مقایسه‌ی مدل پواسون تعمیم‌یافته و مدل دوجمله‌ای منفی در شرایط خاص پرداختند و مدل پواسون تعمیم‌یافته را به‌عنوان آمیخته‌ای از پواسون بیان کردند. لی [۶] مدل پواسون با اثرهای ثابت و تصادفی را در نظر گرفته است. سولیس-تراپالا و فیرول [۸] به بررسی مسئله‌ی بیش‌پراکنش و آزمون‌های مرتبط با آن پرداخته‌اند. معایب استفاده از روش‌های کم‌ترین توان‌های دوم خطی و غیر خطی برای برازش داده‌های شمارشی در هال [۳]، کاکس و همکاران [۹] و وینکلمن [۱۰] بیان شده است. تاریخچه‌ای از ویژگی مدل‌های نام‌برده در کامرون و ترایود [۱] و هاردین و هیلبه [۲] آمده است.

ساختار مقاله به‌صورت زیر است. ابتدا در بخش دوم مدل رگرسیونی لگاریتم خطی معرفی می‌شود. در ادامه، مدل پواسون معرفی و مسئله‌ی بیش‌پراکنش و آزمون‌های مرتبط با آن مطرح می‌شوند. سپس، مدل دوجمله‌ای منفی و مدل پواسون تعمیم‌یافته به‌عنوان مدل‌های جایگزین برای مدل پواسون بررسی می‌گردند. در بخش سوم به مقایسه‌ی مدل‌های رگرسیونی برای داده‌های شمارشی با استفاده از رسم نمودار می‌پردازیم. همچنین در بخش چهارم مدل مناسب برای بررسی اثر اجزای سندرم متابولیک بر فاکتور WBC[1] در کودکان معرفی می‌گردد. داده‌های مورد بررسی طی سال‌های ۱۳۸۷- ۱۳۸۶ از بین افراد چاق ۶ تا ۱۲ ساله در شهر اصفهان به‌منظور تعیین عوامل مؤثر بر تعداد گلبول‌های سفید خون جمع‌آوری شده‌اند.

۲- مدل‌های رگرسیونی متداول برای داده‌های شمارشی

۲-۱- مدل رگرسیون لگاریتم خطی

مدل لگاریتم‌خطی به‌صورت $y_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i)$ معرفی می‌شود که در آن توزیع اثرهای تصادفی را نرمال $u_i \overset{iid}{\sim} N(0, \sigma^2)$ فرض‌می‌کنیم. بنا بر این، متغیرهای پاسخ y_i $(i = 1, \ldots, n)$ دارای توزیع لگ-نرمال با میانگین شرطی $E(Y_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \frac{1}{2}\sigma^2)$ و واریانس $\mathrm{var}(Y_i) = \phi E^2(Y_i)$ می‌باشند که در آن $\phi = e^{\sigma^2} - 1$. استفاده از تابع لگاریتم برای متغیر پاسخ در این مدل منجر به عدم پذیرش

مقدار صفر برای متغیر پاسخ می‌شود. یک راه حل افزودن مقدار ثابتی به متغیر پاسخ است. بنا بر این مدل به $\log(y_i + c) = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$ تبدیل می‌شود که درآن $\log(y_i)$ تابع غیر خطی از x_i می‌باشد. راه حل دیگر حذف مقادیر صفر از مشاهدات است که در این صورت اطلاعات زیادی نادیده گرفته می‌شوند. از معایب این مدل اریب بودن براورد پارامترها است ([۱]، [۹] و [۱۰]).

۲-۲- پواسون

ساختار مدل پواسون به‌عنوان عضوی از خانواده‌ی نمایی یک پارامتری به‌صورت زیر است. فرض کنید که متغیرهای پاسخ Y_i مستقل بوده و دارای توزیع پواسون با میانگین‌های μ_i برای $i = ۱,\ldots,n$ باشند، که در آن $\eta_i = \log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$، $\boldsymbol{\beta}$ و $\boldsymbol{x}_i$ به‌ترتیب بردارهای $(k \times ۱)$ بعدی از پارامترها و متغیرهای توضیحی هستند. تابع درستنمایی به‌صورت زیر است:

$$\log L(\boldsymbol{\beta}) = \sum_{i=۱}^{n}\left\{-\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) + y_i\boldsymbol{x}_i'\boldsymbol{\beta} - \log(y_i!)\right\}. \qquad (۱)$$

محاسبه‌ی براورد پارامترها مستلزم حل معادلات غیر خطی $\sum_{i=۱}^{n}\{y_i - \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\}\boldsymbol{x}_i = ۰$ توسط روش کم‌ترین توان‌های دوم تکراری وزنی است می‌توان نشان داد که براورد پارامتر $\boldsymbol{\beta}$ توسط معادله‌ی $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-۱}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}^*)$ به‌صورت روش عددی- تکراری صورت می‌گیرد که در آن ماتریس وزنی $\boldsymbol{W}$ قطری با اعضای

$$w_{ii} = \frac{۱}{\mathrm{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^۲$$

است و $\boldsymbol{y}^* = \boldsymbol{x}'\hat{\boldsymbol{\beta}} + (\boldsymbol{y} - \boldsymbol{\mu})\frac{\partial\eta}{\partial\mu}$ بردار پاسخ اصلاح‌شده است. براورد پارامترها، تحت شرایط نظم، به‌طور مجانبی دارای توزیع نرمال با میانگین $\boldsymbol{\beta}$ و ماتریس واریانس- کوواریانس $\boldsymbol{A}^{-۱}$ است. به‌عبارت دیگر $\sqrt{n}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}) \xrightarrow{d} N(۰, \boldsymbol{A}^{-۱})$ که در آن

$\boldsymbol{A} = \lim_{n\to\infty} E\left[\frac{1}{n}\sum_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i\boldsymbol{x}_i'\right]$. علاوه بر این ویژگی، سازگار بودن براورد پارامترها نیز برقرار است [۱].

در مسائل کاربردی ممکن است با داده‌های شمارشی روبرو شویم که براورد واریانس مشاهده‌شده از واریانس مورد انتظار بیش‌تر باشد. این ویژگی، بیش‌پراکنش نامیده می‌شود. مدل پواسون برای برازش داده‌های شمارشی با مسئله‌ی بیش‌پراکنش مناسب نیست.

۳-۲- دوجمله‌ای منفی

متداول‌ترین مدل برای برازش داده‌های شمارشی با ویژگی بیش‌پراکنش، دوجمله‌ای‌منفی است که به دو صورت نوع اول (NB-۱) و نوع دوم (NB-۲) بیان می‌شوند و به‌ترتیب برای تحلیل داده‌ها با مسئله‌ی بیش‌پراکنش ثابت و متغیر به کار می‌روند ([۱] و [۱۰]).

۳-۲-۱- بیش‌پراکنش ثابت

فرض کنید که برای $i = 1,\ldots,n$ متغیرهای پاسخ Y_i به شرط λ_i ها دارای توزیع پواسون با میانگین λ_i باشند که در آن λ_i ها متغیرهای تصادفی با توزیع $\Gamma(\mu_i,\delta)$ هستند که $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + offset_i)$. در این صورت مدل آمیخته‌ی پواسون- گاما برای چگالی حاشیه‌ای y_i به ازای $y_i = 0,1,2,\ldots$ به‌صورت $NB\left(\mu_i, \frac{\delta}{\delta+1}\right)$ می‌باشد که درآن $\frac{\mathrm{var}(Y_i)}{E(Y_i)} = \frac{1+\delta}{\delta}$ ساده‌ترین معیار برای بیان بیش‌پراکنش در مدل است. به‌دلیل یکسان بودن این نسبت برای همه‌ی مشاهدات به آن بیش‌پراکنش ثابت می‌گویند. اگر $\alpha = 1/\delta$ آنگاه $\mathrm{var}(Y_i) = (1+\alpha)E(Y_i)$ که نشان‌دهنده‌ی وجود رابطه‌ای خطی بین واریانس و میانگین می‌باشد. اگر $\alpha = 0$ آنگاه مدل به پواسون تبدیل می‌شود.

۲-۳-۲- بیش‌پراکنش متغیر

مدل دوجمله‌ای منفی می‌تواند به‌صورت ترکیبی از مدل پواسون با میانگین λ_i و اثر تصادفی واحدها معرفی شود که در آن با فرض تابع پیوند لگاریتم داریم $\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i = \log(\mu_i) + \log(u_i)$ و u_iها دارای توزیع گاما با میانگین یک، به‌عنوان مثال $\text{Gamma}(v, v)$ باشند. بعد از انجام محاسبات جبری، چگالی کناری Y به‌صورت زیر به دست می‌آید [۱۰]:

$$f(y_i \mid \boldsymbol{x}_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\mu_i\alpha}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\mu_i\alpha}\right)^{y_i} \qquad (۲)$$

که در آن $\alpha = 1/v$ ، $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + offset_i)$ میانگین و $\mu_i(1+\alpha\mu_i)$ واریانس مدل دوجمله‌ای منفی هستند. بنا بر این میزان بیش‌پراکنش برابر است با $(1+\alpha\mu_i)$ که اگر $\alpha = 0$ آنگاه به مدل پواسون تبدیل می‌شود. با توجه به این‌که میزان بیش‌پراکنش به واحدها بستگی دارد به آن بیش‌پراکنش متغیر گویند.

محاسبه‌ی براورد توأم پارامترهای α و $\boldsymbol{\beta}$ به‌روش ماکسیمم درستنمایی مستلزم حل معادلات زیر است:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{1+\alpha\mu_i} x_i = 0 \qquad (۳)$$

$$\sum_{i=1}^{n} \left[\frac{1}{\alpha^2} \left\{ \log(1+\alpha\mu_i) - \sum_{j=1}^{y_i} \frac{1}{\alpha^{-1} + j - 1} \right\} - \frac{(y_i + \alpha^{-1})\mu_i}{1+\alpha^{-1}\mu_i} + \frac{y_i}{\alpha} \right] = 0$$

(۴)

معادلات غیر خطی پیچیده‌ی بالا توسط روش‌های عددی تکراری مانند نیوتن-رافسن قابل حل است که این عموماً توسط نرم‌افزار های آماری مانند STATA، SAS و R محاسبه می‌شوند.

۳-۳-۲- آزمون‌های بیش‌پراکنش

آزمون‌های متفاوتی برای تشخیص بیش‌پراکنش در مدل وجود دارند که در زیر به چند نوع آن اشاره می‌شود [۳]:

۱. لگاریتم درستنمایی مدل‌های پواسون و دوجمله‌ای منفی در آزمون نسبت درستنمایی (LR) برای نشان دادن بیش‌پراکنش به کار برده می‌شود که در آن فرضیه‌های آماری به‌صورت $H_0: E(Y_i) = \mathrm{var}(Y_i)$ و $H_1: E(Y_i) < \mathrm{var}(Y_i)$ بوده و $-2\log(\mathrm{LR})$ از توزیع χ^2_1 پیروی می‌کند.

۲. با توجه به این‌که واریانس دوجمله‌ای منفی $V(\mu,\phi) = \mu + \phi\mu^2$ بوده که در آن μ میانگین مدل است، کافی است که آزمون $H_0: \phi = 0$ را در برابر $H_1: \phi \neq 0$ انجام دهیم. در صورتی که فرض صفر رد نشود واریانس با میانگین برابر بوده و در مدل بیش‌پراکنش وجود ندارد. آماره آزمون برابر است با

$$T_1^2 = \frac{\left[\sum_{i=1}^{n}\left\{(y_i - \hat{\mu}_i)^2 - (1-\hat{\mu}_i)\hat{\mu}_i\right\}\right]^2}{2\sum_{i=1}^{n}\hat{\mu}_i^2} \qquad (۵)$$

که بطور مجانبی دارای توزیع χ^2_1 است.

۳. اگر واریانس مدل پواسون را به‌صورت $V(\mu,\phi) = \phi\mu$ نمایش دهیم که در آن μ میانگین مدل است آنگاه کافی است آزمون $H_0: \phi = 1$ را در برابر $H_1: \phi = 1$ انجام دهیم. آماره‌ی آزمون به‌صورت

$$T_2^2 = \frac{\left[\sum_{i=1}^{n}\left\{(y_i - \hat{\mu}_i)^2 - (1-\hat{\mu}_i)\hat{\mu}_i\right\}\right]^2}{2n\hat{\mu}_i} \qquad (۶)$$

است که به‌طور مجانبی دارای توزیع χ^2_1 است.

بیش‌پراکنش به دو صورت ذاتی و آشکار است که برای تعدیل هر یک روش‌هایی وجود دارد. بنا بر این منطقی است که قبل از بیان میزان بیش‌پراکنش، ابتدا ذاتی یا آشکار بودن آن مشخص شود. بیش‌پراکنش آشکار ممکن است به‌دلایل زیر به وجود آید:

آ) متغیرهای توضیحی مهم وارد مدل نشده باشند؛
ب) مشاهدات شامل مقادیر پرت باشند؛
ج) جملات اثر متقابل مناسب در مدل وجود نداشته باشد؛
د) تبدیل نامناسب برای متغیر توضیحی استفاده شده باشد.

برای اصلاح بیش‌پراکنش آشکار یک روش آن است که اثرات متقابل و متغیرهای توضیحی مورد نیاز را وارد مدل نموده، اثر مشاهدات دور افتاده را کنترل کرده و تبدیلات لازم را در صورت لزوم انجام داد. اما بیش‌پراکنش ذاتی زمانی رخ می‌دهد که بعضی از فرض‌های اساسی مدل‌برقرار نباشند. بنا بر این با به کار بردن مدلی مناسب‌تر، این نوع بیش‌پراکنش کنترل می‌شود.

۴-۲- پواسون تعمیم‌یافته

مدل مناسب دیگری برای تحلیل داده‌های شمارشی با مسئله‌ی پیش‌پراکنش و یا کم‌پراکنش، پواسون تعمیم‌یافته (GP) است [۴]. فرض کنید که Y_iها $(i=1,\ldots,n)$ متغیرهای تصادفی با تابع جرم احتمال زیر باشند:

$$f(y_i \mid \boldsymbol{x}_i) = \frac{\mu(\boldsymbol{x}_i)\{\mu(\boldsymbol{x}_i)+(\varphi-1)y_i\}^{y_i-1}}{y_i!} \times \varphi^{-y_i}\exp\left[-\frac{\{\mu(\boldsymbol{x}_i)+(\varphi-1)y_i\}}{\varphi}\right] \tag{۷}$$

که در آن φ پارامتر پراکندگی و $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ است که μ_i و $\varphi^2\mu_i$ به‌ترتیب امید ریاضی و واریانس Y_iها هستند. واضح است که اگر $\varphi=1$ آنگاه مدل GP به مدل پواسن بامیانگین μ_i تبدیل می‌شود. اگر $\varphi<1$ آنگاه کم‌پراکنش در مدل رخ می‌دهد و اگر $\varphi>1$ آنگاه با مسئله‌ی بیش‌پراکنش مواجه خواهیم شد. با مشتق‌گیری از لگاریتم تابع درستنمایی و خلاصه کردن محاسبات جبری می‌توان نشان داد که براورد پارامترها مستلزم حل معادلات زیر با استفاده از روش‌های عددی تکراری است:

$$\frac{\partial l}{\partial \varphi}=\sum_{i=1}^{n}\left\{\frac{y_i(y_i-1)}{\mu_i+(\varphi-1)y_i}-\frac{y_i}{\varphi}+\frac{\mu_i-y_i}{\varphi^2}\right\}=0 \qquad (8)$$

$$\frac{\partial l}{\partial \beta r}=\sum_{i=1}^{n}\left\{\frac{1}{\mu_i}-\frac{1}{\varphi}+\frac{y_i-1}{\mu_i+(\varphi-1)y_i}\right\}\frac{\partial \mu_i}{\partial \beta r}=0 \qquad (9)$$

که در آن βr ، r امین ضریب رگرسیونی می‌باشد.

۳- مقایسه‌ی مدل‌های متداول برای داده‌های شمارشی

۳-۱- مدل دوجمله‌ای منفی و پواسون تعمیم‌یافته

برای مقایسه‌ی مدل‌ها ابتدا فرض می‌کنیم که گشتاورهای اول و دوم آن‌ها یکسان باشند. سپس توابع جرم احتمال و نیز میزان چولگی آن‌ها را مقایسه خواهیم کرد. به‌طور کلی دو مدل نام‌برده دارای اختلافات جزئی هستند. اما در حالتی که تعداد صفرها در مشاهدات مورد بررسی زیاد شود اختلاف مدل‌ها بیش‌تر نمایان می‌گردد. به‌علت پیچیدگی بررسی ویژگی‌های توزیع پواسون تعمیم‌یافته و به خصوص تابع مولد گشتاور آن، بیش‌تر مدل دوجمله‌ای منفی در برازش داده‌های بیش‌پراکنش استفاده می‌شود. با توجه به این‌که این دو مدل بیش‌تر در حالت بیش‌پراکنش مقایسه می‌شوند بنا بر این، پارامتر φ درمدل پواسون تعمیم‌یافته مقادیر کم‌تر از یک را اتخاذ نمی‌کند.

هری و رانگ [۴] تفاوت این دو مدل را به‌صورت زیر بیان می‌کنند:

۱. مدل دوجمله‌ای منفی در مشاهده‌ی صفر، احتمال بیش‌تری نسبت به مدل پواسون تعمیم‌یافته دارد.

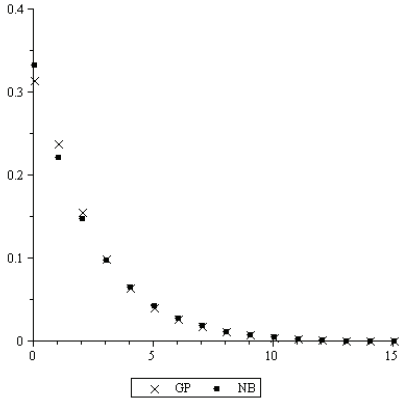
۲. در پیشامد یک اگر میانگین‌مدل‌ها بیش‌تر از مقدار ثابتی باشند آنگاه مدل دوجمله‌ای منفی احتمال بیش‌تری نسبت به مدل پواسون تعمیم‌یافته دارد و اگر میانگین به صفر نزدیک شود آنگاه مدل دوجمله‌ای منفی احتمال کم‌تری نسبت به مدل پواسون تعمیم‌یافته خواهد داشت.

۳. با افزایش مقدار مشاهدات مدل پواسون تعمیم‌یافته دارای دم سنگین‌تری نسبت به مدل دوجمله‌ای منفی است.

هری و رانگ [۴] نتیجه‌ی (۳) را با استفاده از تقریب استرلینگ بیان کرده‌اند اما در این مقاله ما با رسم نمودار دو مدل با میانگین و واریانس یکسان ۲ و ۶، مشاهده کردیم که برای مشاهدات بیش‌تر از ۳ هر دو مدل احتمالات نزدیک به یکدیگر را اختیار می‌کنند و دارای دم‌های یکسان هستند.

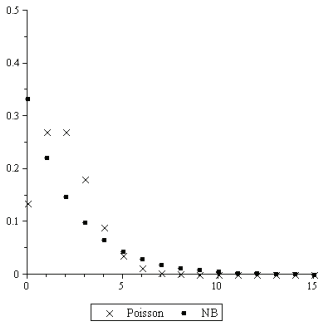
۲-۳- مقایسه‌ی مدل‌های دوجمله‌ای منفی و پواسون تعمیم‌یافته با پواسون

در این قسمت نیز با شرط تساوی دو گشتاور مرتبه‌ی اول مدل‌ها به مقایسه‌ی مدل‌های دوجمله‌ای منفی و پواسون تعمیم‌یافته با پواسون با استفاده از روش‌های گرافیکی می‌پردازیم. لازم به ذکر است که برای همه‌ی مدل‌ها میانگین ۲ و واریانس برحسب میزان پراکنش در مدل فرض شده است. برای مقایسه‌ی آسان‌تر مدل‌ها با پراکنش متفاوت و مدل پواسون از نمودارهای پیوسته استفاده می‌شود. به‌ترتیب در نمودارهای ۲ و ۳ مدل پواسون با مدل‌های پواسون تعمیم‌یافته و دوجمله‌ای منفی مقایسه شده است. همان طور که مشاهده‌می‌شود مدل پواسون در مشاهده‌ی صفر احتمال کم‌تری دارد و با افزایش مقادیر مشاهدات سریع‌تر به سمت صفر همگرا می‌شود. در شکل‌های ۴ و۵ مشاهده می‌شود که با کاهش میزان پراکنش این دو مدل به پواسون نزدیک می‌شوند.

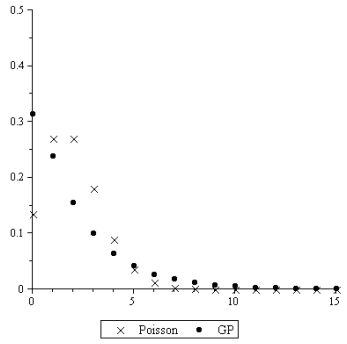


شکل ۱- مقایسه‌ی مدل دوجمله‌ای منفی و پواسون تعمیم‌یافته

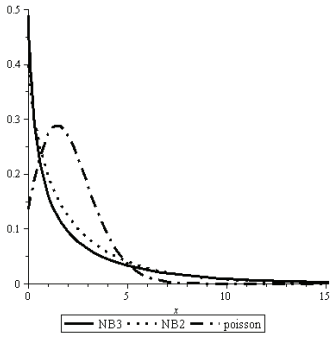
لازم به‌ذکر است برای بیان روشن‌تر اختلاف این مدل‌ها با پواسون و به‌دلیل نزدیک بودن دو نمودار پواسون تعمیم‌یافته و دوجمله‌ای منفی از میزان پراکنش متفاوت استفاده شده است.



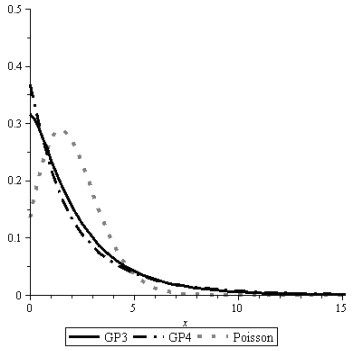
شکل ۲- مقایسه‌ی پواسون‌تعمیم‌یافته و پواسون



شکل ۳- مقایسه‌ی دوجمله‌ای‌منفی و پواسون



شکل ۴- مقایسه‌ی پواسون‌تعمیم‌یافته با پراکنش ۳ و ۴ و پواسون



شکل ۵- مقایسه‌ی دوجمله‌ای‌منفی با پراکنش ۲ و ۳ و پواسون

۴- مثال کاربردی

هدف از انجام این تحقیق علاوه بر معرفی مدل رگرسیونی مناسب، بیان تأثیر هر یک از اجزای سندرم متابولیک بر تعداد گلبول‌های سفید در هر میلی‌متر مکعب (WBC) است. داده‌های مورد بررسی در سال‌های ۱۳۸۷-۱۳۸۶ از بین کودکان چاق ۶ تا ۱۲ شهر اصفهان که به مراکز بهداشت، مراکز درمان چاقی و کلینیک تحقیقاتی سندرم متابولیک از بخش شناسایی امراض قلبی کودکان و مرکز تحقیق قلب و عروق اصفهان (ICRC)[۲] مراجعه می‌کردند، انتخاب شده‌اند. این تحقیق با همکاری سازمان سلامت جهانی و علوم دارویی دانشگاه اصفهان انجام شده است. کودکان با عقب‌ماندگی ذهنی، مسائل دارویی مزمن یا کسانی که از دارو استفاده می‌کنند و کودکانی که دارای چهره‌ی غیرعادی یا دارای سابقه‌ی بیماری عفونی هستند، دو هفته قبل از انجام مطالعه از نمونه خارج شده‌اند. متغیرهای توضیحی که در این تحلیل استفاده می‌شوند عبارت‌اند از: سن (Age)، جنسیت (Sex) که در آن مذکر بودن با یک نشان داده شده است، اندازه‌ی دور کمر (Waist)، اندازه دور ران (Hip)، نسبت کلسترول کل به HDL[۳]، قند خون ناشتا (Fbs) و نسبت تری‌گلیسیرید خون به HDL.

نتایج با استفاده از نرم‌افزارهای SPSS، STATA و SAS به دست آمده‌اند. نتایج جدول ۱ به‌صورت میانگین ± انحراف معیار، فراوانی با فاصله‌ی اطمینان ۹۵ درصد گزارش شده‌اند. برای مقایسه‌ی ویژگی‌های دختران و پسران از آزمون تی مستقل استفاده شده است. نتایج نشان می‌دهد که ۴۷٪ کودکان پسر هستند و میانگین متغیرهای سن، میزان نمایه‌ی توده‌ی بدن، تعداد گلبول‌های سفید خون، نسبت کلسترول به HDL-C[۴]، اندازه‌ی دور باسن، اندازه‌ی دور کمر و نسبت LDL-C[۵] به HDL-C پسران و دختران تفاوت معناداری ندارند. میانگین قند خون ناشتا در پسران بیش‌تر از دختران است و بر اساس آزمون آماری بین دو متغیر جنسیت و قند خون ناشتا اختلاف معناداری وجود دارد ($p > ۰/۰۵$). از نتایج دیگر این پژوهش می‌توان به ارتباط معنادار بین جنسیت و نسبت تری‌گلیسیرید به HDL-C اشاره کرد به‌طوری که میزان نسبت تری‌گلیسیرید به HDL-C در پسران بیش‌تر است ($p < ۰/۰۵$).

جدول ۱- مقایسه‌ی میانگین متغیرها بر اساس جنسیت

	پسر (۱۵۲ نفر)	دختر (۱۴۰ نفر)	p
سن (سال)	۱۰/۱۶ ± ۲/۵۹	۹/۷۸ ± ۲/۸۷	۰/۴۶۰
BMI ($kg/m^۲$)	۲۴/۷۶ ± ۴/۶۶	۲۲/۵۹ ± ۳/۷۸	۰/۲۰۲
تعداد گلبول‌های سفید ($celles/mm^۳$)	۷۳۳۳/۲۹ ± ۲۰۰۸/۳۰	۷۴۱۹/۱۴ ± ۲۰۸۵/۰۳	۰/۷۴۱
کلسترول به HDL-C	۴/۰۸ ± ۱/۰۸	۳/۹۸ ± ۰/۹۷	۰/۳۷۵
قند خون ناشتا (mg/dl)	۸۸/۲۵ ± ۹/۹۴	۸۶/۴۶ ± ۷/۶۷	۰/۰۰۳
تری گلیسیرید به HDL-C	۳/۳۱ ± ۲/۷۱	۲/۹۶ ± ۱/۸۷	۰/۰۲۳
اندازه‌ی دور باسن (cm)	۹۴/۱۸ ± ۱۱/۰۹	۸۹/۷۶ ± ۱۰/۵۷	۰/۷۹۲
اندازه‌ی دور کمر (cm)	۸۴/۷۲ ± ۱۳/۲۴	۷۹/۸۰ ± ۱۲/۴۰	۰/۹۶۳
LDL-C به HDL-C	۲/۴۴ ± ۰/۹۰	۲/۴۰ ± ۰/۷۶	۰/۲۷۰

معیارهای انتخاب مدل در جدول ۳ آمده است. مشاهده می‌شود که مدل پواسون دارای مقادیر معیار اطلاع آکائیکی (AIC) و معیار اطلاع بیزی (BIC) بزرگی است که دلیل بر نامناسب بودن مدل می‌باشد. اما در سایر مدل‌ها معیارهای AIC و BIC کاهش یافته است. با توجه به این که این مقادیر برای مدل پواسون تعمیم‌یافته کم‌تر است، لذا به‌عنوان مدل برازش‌داده‌شده‌ی مناسب انتخاب می‌شود. جدول ۲، براورد پارامترها و انحراف استاندارد آن‌ها را نشان می‌دهد و در صورتی که مقادیر بیان‌شده از $z_{۰/۹۷۵} = ۱/۹۶$ بیش‌تر و یا از $z_{۰/۰۲۵} = -۱/۹۶$ کم‌تر باشند دارای اثر معناداری در مدل هستند. با توجه به وجود مسئله‌ی بیش‌پراکنش در مدل پواسون، انحراف استاندارد ضرایب رگرسیونی کم‌تر از مقدار واقعی براورد می‌شوند. در مدل پواسون تعمیم‌یافته رابطه‌ی متغیر نسبت کلسترول به HDL-C با تعداد گلبول‌های سفید خون ارتباطی منفی و معنادار دارد که با افزایش آن، از میانگین تعداد گلبول‌های سفید خون در میلی‌متر مکعب کاسته می‌شود و رابطه‌ی متغیر نمایه‌ی توده‌ی بدن با تعداد گلبول‌های سفید خون ارتباطی مثبت و معنادار است که با افزایش آن، به میانگین تعداد گلبول‌های سفید خون در میلی‌متر مکعب افزوده می‌شود.

جدول ۲- براورد پارامترها و انحراف استاندارد آن‌ها برای داده‌های تعداد گلبول سفید خون

	مدل رگرسیونی لگاریتم خطی	پواسون	دوجمله‌ای منفی	پواسون تعمیم‌یافته
سن	-۰/۰۷۹(۰/۰۲۰)	-۰/۰۱۵(۰/۰۰۰۴)	-۰/۰۱۳(۰/۰۰۷)	-۰/۰۰۸(۰/۰۰۷)
نمایه‌ی توده‌ی بدن	-۰/۰۵۸(۰/۰۱۷)	۰/۰۲۳(۰/۰۰۰۳)	۰/۰۲۶(۰/۰۰۷)	۰/۰۲۱(۰/۰۰۵)
جنسیت	-۰/۱۵۱(۰/۰۸۰)	-۰/۰۵۰(۰/۰۰۱)	-۰/۰۴۶(۰/۰۳۰)	-۰/۰۴۷(۰/۰۲۹)
کلسترول به HDL-C	۰/۳۱۲(۰/۱۰۵)	-۰/۰۶۸(۰/۰۰۲)	-۰/۰۶۷(۰/۰۴۱)	-۰/۰۵۰ (۰/۰۳۷)
قند خون ناشتا	۰/۰۴۸(۰/۰۰۳)	۰/۰۰۳(۰/۰۰۰۱)	۰/۰۰۳(۰/۰۰۲)	۰/۰۰۳(۰/۰۰۲)
تری گلیسیرید به HDL-C	-۰/۰۶۵(۰/۰۲۴)	۰/۰۳۴(۰/۰۰۰۴)	۰/۰۳۳(۰/۰۰۹)	۰/۰۳۰(۰/۰۰۸)
اندازه‌ی دور کمر	۰/۰۰۲(۰/۰۰۵)	۰/۰۰۰۲(۰/۰۰۰۱)	۰/۰۰۰۱(۰/۰۰۲)	۰/۰۰۰۳(۰/۰۰۲)
اندازه‌ی دور باسن	۰/۰۶۵(۰/۰۰۸)	-۰/۰۰۵(۰/۰۰۰۲)	-۰/۰۰۶(۰/۰۰۳)	-۰/۰۰۵(۰/۰۰۳)
LDL-C به HDL-C	-۰/۱۶۴(۰/۱۱۷)	۰/۰۶۸(۰/۰۰۲)	۰/۰۶۸(۰/۰۴۲)	۰/۵۴۶(۰/۰۳۹)
مقدار ثابت	۰/۲۲۲(۰/۰۱۸)	۸/۷۵۳(۰/۰۰۹)	۸/۷۷۴(۰/۲۰۲)	۸/۶۸۰(۰/۱۸۷)
پارامتر مقیاس	۰/۴۴۵(۰/۰۳۶)	ــــ	۰/۰۶۰(۰/۰۰۵)	۰/۹۵۳(۰/۰۰۲)

جدول ۳- معیارهای انتخاب مدل

	مدل رگرسیونی لگاریتم خطی	پواسون	دوجمله‌ی منفی	پواسون تعمیم‌یافته
AIC	۵۷۹۲/۹	۱۳۷۳۸۹	۵۲۱۶	۵۲۰۹/۴
BIC	۵۸۲۹/۶	۱۳۷۴۶۲	۵۲۵۶	۵۲۴۹/۸

۵- نتیجه‌گیری

با توجه به مشکلاتی که برازش مدل پواسون برای داده‌های شمارشی با مسئله‌ی بیش‌پراکنش دارد، در این مقاله مدل‌های جایگزین متفاوتی پیشنهاد شد و با استفاده از نمودارها، ویژگی‌های آن‌ها مورد بررسی قرار گرفت. در ادامه با بیان مثالی کاربردی،

مدل‌های مختلف به داده‌ها برازش داده شد و با استفاده از معیارهای AIC و BIC مدل پواسون تعمیم‌یافته، که دو موضوع بیش‌پراکنش و کم‌پراکنش را در تحلیل داده‌ها منظور می‌کند، نسبت به بقیه به‌عنوان مدل مناسب انتخاب شد. یافته‌های مطالعه‌ی حاضر که در اصفهان و بر روی کودکان چاق ۱۲-۶ ساله انجام شد حاکی از آن است که بین تعدادی از اجزای سندرم متابولیک و میانگین تعداد گلبول‌های سفید خون در هر میلی‌متر مکعب ارتباط برقرار است. بازبینی پیشینه‌ی تحقیق نشان می‌دهد که تا کنون هیچ مقاله‌ای در مورد ارتباط اجزای سندرم متابولیک و تعداد گلبول‌های سفید منتشر نشده است. بیش‌تر مطالعات مشاهده‌ای نیز بر ارتباط این سندرم با بیماری‌های کرونری قلبی و شیوع این سندرم تمرکز داشته‌اند [۷]. به‌عنوان مثال، جلالی و همکاران شیوع سندرم متابولیک در بزرگسالان اکبرآباد کوار فارس را در سال ۱۳۸۷ بررسی کرده‌اند [۵]. همچنین از سال ۱۳۸۲ تاکنون تقریباً ۴۰ کار آزمایی بالینی منتشر شده است که اثر متابولیک دو یا تعدادی بیش‌تر از رژیم‌های متفاوت از نظر کربوهیدرات دریافتی را در بزرگسالان دارای ویژگی سندرم متابولیک بررسی کرده‌اند [۷]. در مطالعه‌ی حاضر برخلاف بیش‌تر مطالعه‌های قبلی به بررسی گروه سنی کودکان پرداخته شده است و با توجه به اهمیت سیستم دفاعی بدن به معرفی مدل‌های شمارشی مختلف به‌منظور تعیین اثر اجزای سندرم متابولیک بر تعداد گلبول‌های سفید پرداخته است.

سپاس‌گزاری

از تحصیلات تکمیلی دانشگاه اصفهان به خاطر فراهم کردن تسهیلات پژوهشی قدردانی می‌شود. همچنین از خانم دکتر رؤیا کلیشادی عضو هیئت‌علمی دانشگا علوم پزشکی اصفهان جهت همکاری مستمر و از هیئت داوران نیز به خاطر پیشنهادهای سازنده سپاس‌گزاری می‌شود.

توضیحات

1. White Blood Cell
2. Isfahan Cardiovascular Research Center
3. High Density Lipoprotein

4. High Density Lipoprotein Cholesterol
5. Low Density Lipoprotein Cholesterol

مرجع‌ها

[1] Cameron, A.C. and Trivedi, P.K. (1998). Regression analysis of count data. *Econometric Society Monogrphs*.

[2] Hardin, J.W. and Hilbe J.M. (2007). *Generalized Linear Models and Extensions*. Stata Press. College Station, Texas.

[3] Hall, D.B. (2000). Zero inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**, 1030-1039.

[4] Harry, J. and Rong, Z. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal*, **47**, 219-229.

[5] Jalali, R.; Vasheghani, M.; Dabbaghmanesh, M.H. and Ranjbaromrani, GH. (2009). Prevalence of metabolic syndrome among adults in a rural area. *Iranian Journal of Endocrinology and Metabolism*, **11**, 405-414.

[6] Lee, Y. (2002). Fixed effect versus random effect models for evaluating therapeutic preferences. *Statistics in Medicine*, **21**, 2325-2330.

[7] Mirmiran, P.; Saidpour, A.; Hosseinpour Niazi, S. and Azizi, F. (2010). Glycemic index and glycemic load in relation with metabolic syndrome and its components. *Iranian Journal of Endocrinology and Metabolism*, **11**, 615-625.

[8] Solis-Trapala, I.L. and Farewell, V.T. (2005). Regression analysis of over-dispersion correlated count data with subject specific covariates. *Statistics in Medicine*, **24**, 2557-2575.

[9] Coxe, S.; West, S.G. and Aiken, L.S. (2009). The analysis of count data: a gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, **91**, 121–136.

[10] Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer-Verlag, New York.

فاطمه حسن‌زاده
فوق لیسانس آمار
اصفهان، دانشگاه اصفهان، گروه آمار.
رایانشانی: f_hassanzadeh65@yahoo.com

ایرج کاظمی
دکتری آمار
اصفهان، دانشگاه اصفهان، گروه آمار.
رایانشانی: i.kazemi@stat.ui.ac.ir