تأثیر تغییر توزیع داده‌ها در آمیزه‌ای متناهی از مدل‌های خطی تعمیم‌یافته نیم پارامتری

فرزاد اسکندری؛ دانشگاه علامه طباطبایی، دانشکدهٔ علوم ریاضی و رایانه

دریافت ۹۷/۰۲/۲۱ پذیرش ۹۷/۰۶/۲۶

چکیده

انتخاب مدل آماری مناسب برای متغیر پاسخ یکی از مهم‌ترین مسایل در آمیزه‌ای متناهی از مدل‌های خطی تعمیم یافته است. یکی از توزیع‌هایی که در آمیزه‌ای متناهی از مدل‌های خطی تعمیم یافته نیم پارامتری با مشکل روبه‌رو است، توزیع پواسون است. در این مقاله، برای غلبه بر بار محاسباتی و بیش پراکندگی، آمیزه‌ای متناهی از مدل‌های خطی تعمیم-یافته نیم پارامتری با استفاده از توزیع دوجمله‌ای منفی (GFMMNB) به‌جای آمیزه‌ای متناهی از مدل‌های خطی تعمیم یافته نیم پارامتری با استفاده از توزیع پواسون (GFMMP) پیشنهاد می‌شود. کارایی مدل‌های آمیزه‌ای متناهی از مدل‌های خطی تعمیم یافته نیم پارامتری با استفاده از توزیع دوجمله‌ای منفی نسبت مدل‌های آمیزه‌ای متناهی از مدل‌های خطی تعمیم یافته نیم پارامتری با استفاده از توزیع پواسون برای داده‌های شبیه‌سازی شده و داده‌های واقعی و با استفاده از میانگین وزنی توان دوم انحرافات (WGMSE) نشان داده شده است.

واژه‌های کلیدی: مدل‌های آمیزه، الگوریتم EM، تابع تاوان، هم‌ترازسازی،

مقدمه

مدل‌سازی و تعیین ارتباط بین متغیر پاسخ و متغیرهای کمکی یکی از مهم‌ترین فن‌های آماری است که مورد توجه آمار شناسان و پژوهش‌گران حوزه‌های مختلف علوم کاربردی قرار دارد. انتخاب متغیرهای مؤثر نیز بخشی از مدل‌سازی است که نقشی اساسی در تفسیر درست داده‌ها و استخراج اطلاعات از آن‌ها برای رسیدن به دانش مطلوب و مدیریت دانش داده‌ها فراهم می‌آورد. امروزه به‌دلیل وجود حجم وسیعی از داده‌های پزشکی، مهندسی و محیطی نیاز به مدل‌های آماری انعطاف‌پذیر همانند مدل‌های رگرسیونی نیم‌پارامتری افزایش یافته است. چون با استفاده از این مدل‌ها می‌توان به تحلیل داده‌های پیچیده‌ای پرداخت که در حال توسعه هستند، به‌ویژه زمانی که داده‌ها برگرفته از یک جامعه ناهمگن باشند. بنابراین بعضی مواقع داده‌هایی وجود دارند که همگن و یک‌دست نیستند، در چنین شرایطی آمیزه‌ای متناهی از مدل‌های آماری به‌عنوان ابزاری انعطاف‌پذیر برای مدل‌بندی این‌گونه داده‌ها استفاده می‌شود. این موضوع در پژوهش‌های مختلفی بررسی شده است. یکی از توزیع‌های معروف که معمولاً در تئوری به‌عنوان متغیر پاسخ برای تحلیل داده‌ها در نظر گرفته می‌شود، توزیع پواسون است. یکسان بودن میانگین و واریانس در تئوری و یکسان نبودن در کاربرد می‌تواند به‌عنوان نقطهٔ ضعف این توزیع اعلام شود. زمانی که تعداد متغیرهای کمکی زیاد باشد و آمیزه‌ای متناهی از توزیع‌ها برای متغیر پاسخ در نظر گرفته شود، استفاده از آمیزه‌ای از توزیع‌های پواسون به‌دلیل

*نویسنده مسئول askandari@atu.ac.ir

یکسان نبودن میانگین و واریانس در کاربرد می‌تواند مشکلاتی که از این منظر برای تحلیل داده‌ها ایجاد می‌کند را بسیار افزایش دهد.

در آمیزهٔ متناهی از توزیع‌ها برای وقتی که بیش پراکنش به‌وجود می‌آید، یکی از توزیع‌های جای‌گزین به‌جای توزیع پواسون، توزیع دوجمله‌ای منفی‌است. این موضوع برای وقتی که ساختار داده‌ها در شرایط آمیزه‌ای متناهی از توزیع‌ها برای متغیر پاسخ در نظر گرفته می شود می‌تواند به‌عنوان یک موضوع تحقیقی مناسب باشد که در این مقاله به‌آن در دو حوزه نظری و کاربردی پرداخته شده است. آمیزه متناهی از توزیع‌ها در حالتی که توزیع متغیرهای پاسخ پواسون است در کارهای قبلی بررسی شده است، اما در این مقاله هدف بررسی این موضوع است. چنان‌چه تغییر توزیع را انجام دهیم میزان تغییرات در مباحث نظری و تحلیل داده‌ها چه میزان است؟

لازم به ذکر است استفاده از روش‌های معتبر مانند توابع تاوان در انتخاب متغیرها باعث افزایش دقت در نتیجه‌گیری، کاهش اریبی و کارایی برآورد کننده‌ها می‌شود. به‌منظور کاستن از میزان اریبی‌های ممکن مدل‌سازی، معمولاً تعداد زیادی از متغیرهای کمکی در مراحل اولیهٔ مدل‌سازی از مدل حذف می‌شوند. از سوی دیگر، به‌منظور افزایش قابلیت پیش‌بینی و انتخاب متغیرهای معنی‌دار، آمارشناسان اغلب از حذف‌گام به‌گام و انتخاب بهترین زیرمجموعه متغیرهای کمکی استفاده می‌کنند. این موضوع در این مقاله براساس تغییر توزیع پاسخ در دو حالت مورد نقد و بررسی قرار می‌گیرد.

آن‌چه که باعث انجام پژوهش حاضر شده است موضوع ادغام آمیزه متناهی از مدل‌های آماری به‌همراه مفهوم هم‌ترازسازی است. در این رابطه بر اساس آیین‌نامهٔ مصوب وزارت علوم تحقیقات و فناوری مقرر شد پذیرش دانشجو برای مقطع دکتری بر اساس آزمون به‌صورت عمومی به‌وسیلهٔ سازمان سنجش آموزش کشور برگزار شود، چند برابر ظرفیت پذیرش به دانشگاه‌ها و مؤسسات آموزش عالی معرفی تا براساس مصاحبهٔ علمی نمرهٔ داوطلب به سازمان سنجش برای پذیرش نهایی داوطلبان ارسال شود. هم‌چنین طبق قوانین مقرر شده است به‌طور مستمر هر شش ماه یک بار این آزمون برگزار شده و داوطلبانی که حد نصاب لازم را دارا باشند برای انجام مصاحبه و مانند آن به دانشگاه‌ها معرفی شوند. از آن‌جاکه در یک دورهٔ مشخص که افراد برای مصاحبه به دانشگاه‌ها مراجعه می‌کنند کسانی هستند که در آزمون‌های مختلف سازمان سنجش شرکت کرده‌اند، از این رو، توزیع امتیازها برای آنان متفاوت است و در نتیجه می‌توان اعلام کرد که آمیزه‌ای متناهی از چند جامعه وجود دارد. یکسان‌سازی و یا هم‌ترازسازی زمانی رخ می‌دهد که بتوانیم ترکیبی از توزیع‌ها را داشته باشیم. در این‌جا نکته قابل توجه که باید در نظر گرفت این است که تعداد سئوالاتی (نمونه‌هایی) که فرد باید به آن‌ها پاسخ درست بدهد تا حد نصاب لازم را برای معرفی به‌منظور مصاحبه بیاورد متغیری تصادفی است، که به‌دلیل دو وضعیتی بودن پاسخ (درست-غلط)، می‌تواند از توزیع دوجمله‌ای منفی تبعیت کند. در واقع در اینجا آمیزه‌ای متناهی از مدل‌های آماری با پاسخ دوجمله‌ای منفی را داریم که لازم است مبانی نظری و شیوه برآورد پارامترها و آزمون‌های مربوط به تأثیر و عدم تأثیر عوامل کمکی را بر پاسخ بررسی می‌کنیم. در این بررسی استفاده از توابع تاوان نیز برای افزایش دقت به‌عنوان ایده‌ای جدید می‌تواند استفاده شود. برای این منظور ابتدا به پیشینه‌ای از کارهایی که در زمینهٔ آمیزه متناهی از مدل‌های آماری انجام شده است می‌پردازیم، سپس به کارهایی که در زمینهٔ هم‌ترازسازی انجام شده است می‌پردازیم.

اگرچه کاربرد این روش‌ها در عمل مفید واقع می‌شوند، اما خطاهای تصادفی ذاتی را در مرحلهٔ انتخاب متغیرها نادیده می‌گیرند. بنابراین خواص نظری آن‌ها تقریباً پیچیدگی خاصی دارد که پژوهش‌گر باید به آن توجه کند. علاوه بر این، انتخاب بهترین زیرمجموعه متغیرها دارای ویژگی‌های خاص دیگری است که مهم‌ترین آن‌ها توجه به رعایت پایداری است. موضوع زمان بر بودن و توجه به مباحث محاسبه‌ای این روش‌ها نیز موضوع دیگری است که باید به آن اشاره کرد. اما نکته قابل توجه در انتخاب متغیر تشخیص درست مدل پیش‌نهادی برای رسیدن به جواب واقعی است که در پژوهش حاضر بسیار به آن توجه می‌شود. انجام تحقیقات و رسیدن به جواب مطلوب بر اساس انتخاب متغیرهای کمکی مؤثر مفهومی است که اگر به درستی انجام شود برای پژوهش‌گر این امکان را فراهم می‌کند تا درک بهتری از تغییرات موجود در محیط اطراف خود داشته باشد. این موضوع در شرایطی که ساختار متغیرها و ارتباط بین آن‌ها از یک پیچیدگی خاص مانند ساختار نیم‌پارامتری برخوردار باشد بیش‌تر باید مورد توجه قرار گیرد.

آمیزه‌ای متناهی از توزیع‌های آماری

در بسیاری از موارد برای رویدادهای طبیعی می‌توان رهیافتی آماری برمبنای آمیزهٔ متناهی از مدل‌ها را معرفی کرد تا برمبنای آن گسترهٔ وسیعی از پدیده‌های تصادفی را تحلیل و تفسیر کرد. به‌دلیل کارایی این‌گونه توزیع‌ها به‌عنوان شیوه‌ای بسیار انعطاف‌پذیر برای مدل‌بندی، مدل‌های آمیزه‌ای متناهی از نظر کاربردی و نظری توجه بسیاری را طی سال‌های اخیر به‌خود جلب کرده‌اند. این مدل‌ها در زمینهٔ علوم مختلف مانند علوم زیستی، اقتصادی و اجتماعی و هم‌چنین در زمینه‌های متفاوتی از قبیل نجوم، علم ژنتیک، پزشکی، روان‌پزشکی، علوم مهندسی و بازاریابی به‌طور موفقیت‌آمیزی کاربرد دارند.

تعریف ۱. فرض کنید $Y_1, \dots, Y_n$ یک نمونهٔ تصادفی به اندازهٔ n باشد، به‌طوری‌که Y_j یک بردار تصادفی p بعدی با تابع چگالی احتمال $\mathrm{f}(y_j;\theta_i)$ ، برای $(i = 1, \dots, g)$ روی فضای $\mathcal{R}^p$ است. اگر $\mathrm{f}_i(y_j;\theta_i)$ تابع چگالی j امین متغیر در iامین جامعه باشد، در آن صورت آمیزه‌ای متناهی از توابع چگالی متغیر تصادفی Y بدین‌صورت نوشته می‌شود:

$$f(y_j;\Psi) = \sum_{i=1}^{g} \pi_i \, f_i(y_j;\theta_i) \qquad j = 1,2,\dots,p \qquad (۱)$$

Ψ برداری شامل تمام پارامترهای نامعلوم در مدل آمیزه‌ای است و به‌صورت $\Psi = (\pi_1, \dots, \pi_{g-1}, \xi^T)^T$ تعریف می‌شود. $\pi_1, \dots, \pi_g$ کمیت‌هایی نامنفی هستند که به‌عنوان وزن‌ها نامیده ارایه می‌شود و مقادیری را بین صفر و یک اختیار می‌کنند به‌طوری‌که

$$\sum_{i=1}^{g} \pi_i = 1 \qquad (۲)$$

ξ نیز بردار شامل تمام پارامترهای مجهول $\theta_1, \dots, \theta_g$ است.

در خصوص آمیزه متناهی از توزیع‌های آماری، ارمز و اسکندری (۲۰۱۶) فرض کردند متغیر پاسخ در ارتباط با متغیر کمکی از مدل رگرسیونی نیم پارامتری تعمیم‌یافته تبعیت می‌کند. در آن پژوهش فرض شد متغیر پاسخ Y با مقادیر ممکن $Y \subset \mathcal{R}$ و برداری از متغیرهای کمکی به‌صورت (u, x, z) با $x = (x_1, x_2, \dots, x_q)^T$ به‌عنوان

متغیرهای حقیقی و ضرایب بخش ناپارامتری، $z = (z_1, z_2, \dots, z_p)^T$ ضرایب بخش پارامتری مدل و u یک متغیر تکی است. براساس آن آمیزه‌ای متناهی از مدل رگرسیونی نیم‌پارامتری به صورت زیر تعریف می‌گردد

تعریف ۲. فرض کنید $G = \{f(y;\theta,\phi); (\theta,\phi) \in \Theta \times (0,\infty)\}$ یک خانواده از توابع چگالی پارامتری Y باشد، که در آن $\Theta \subset \mathcal{R}$ و ϕ پارامتر پراکندگی است. گفته می‌شود (u,x,z,Y) آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری با مرتبه K است، هرگاه تابع چگالی شرطی Y به شرط (u,x,z) به‌صورت:

$$f(y;u,x,z,\Psi) = \sum_{k=1}^{K} \pi_k f(y;\theta_k(u,x,z),\phi_k) \qquad (۳)$$

با این شرایط ارایه شود:

الف) $\theta_k(u,x,z) = h(x^T\alpha_k(u) + z^T\beta_k)$ باشد

ب) $\alpha(.)$ برداری شامل تابع‌های نامعلوم از ضرایب رگرسیونی هموار باشد.

ج) بردار پارامتری Ψ بدین‌صورت باشد:

$$\Psi = (\alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_K, \phi, \pi)$$

شامل $\alpha_k = (\alpha_{k1}, \dots, \alpha_{kq})^T$، $\beta_k = (\beta_{k1}, \dots, \beta_{kp})^T$، $\phi = (\phi_1, \dots, \phi_K)^T$ و $\pi = (\pi_1, \dots, \pi_{K-1})^T$.

آمیزه‌ای متناهی از مدل‌های نیم پارامتری ارائه شده به‌وسیلهٔ ارمز و اسکندری (۲۰۱۶) که عضو خانواده نمایی هستند، در تعریف ۲ روشی معمول به‌منظور مدل‌بندی این‌گونه رابطه‌های ناهمگن نامشهود فراهم می‌کند.

اما یک مشکل بزرگ در ساختار چنین خانواده این است که هنگامی که هم‌خطی رخ می‌دهد، برآوردهای حاصل دچار ناپایداری می‌شوند یعنی برمبنای نمونه‌های متفاوت، مقدار برآوردها به‌میزان چشم‌گیری تغییر می‌یابد. برای حل چنین مشکلی می‌توان با صفر کردن یا انقباض بعضی از ضرایب برآوردگرها، اریبی را کمی افزایش ولی به تبع آن واریانس را کاهش داد، در نتیجه میزان دقت پیش‌بینی کل را بهبود می‌بخشد. برای حل چنین مشکلی شیوهٔ انتخاب متغیر تاوانیده را می‌توان مطرح کرد که با ایجاد محدودیت روی مجموع توان دوم مانده‌ها، عمل برآورد پارامترها انجام می‌پذیرد. تیب‌شیرانی (۱۹۹۶) تابع تاوانی با نام کم‌ترین انقباض مطلق و عملگر گزینش (LASSO) و فن و لی (۲۰۰۱) تابع تاوان انحراف مطلق به‌طور هموار بریده شده (SCAD) را معرفی کردند، به‌طوری‌که در حضور این تابع‌های تاوان عمل انتخاب متغیر و برآورد پارامترها به‌طور هم‌زمان انجام می‌پذیرد. بر مبنای فن و لی (۲۰۰۱) یک تابع تاوان خوب تابعی است که این سه ویژگی را دارد:

۱. نااریبی: به‌منظور جلوگیری از اریبی، مقدار برآورد گر باید به مقادیر بزرگ پارامتر نامعلوم نزدیک باشد.

۲. تنکی: برآورد حاصل باید از یک قاعده آستانه‌پذیری تبعیت کند، به این معنی که به‌طور خودکار پارامترهای برآورد شده با مقدار کم را صفر در نظر بگیرد تا از پیچیدگی مدل کاسته شود.

۳. پیوستگی: برآوردهای حاصل از یک تابع تاوان باید پیوسته باشند تا از ناپایداری در پیش‌بینی مدل جلوگیری شود.

بر مبنای فن و لی (۲۰۰۱) تمام تابع‌های تاوان به‌جز تاوان SCAD در برقراری هم‌زمان این سه شرط ناتوان هستند. تابع تاوان SCAD نسبت به روش‌های سنتی انتخاب متغیر هزینه محاسباتی کم‌تری دارد و نسبت به دیگر روش‌های تاوانیده جواب‌های پیوسته ایجاد می‌کند، در نتیجه از تغییرپذیری غیرضروری در مدل اجتناب می‌شود. سانترلی و همکاران (۲۰۱۶) مدل کانوی- ماکسول-پواسون (CMP) را به‌صورت آمیزه‌ای استفاده کرده‌اند و برمبنای

توزیع پواسون به تحلیل داده‌های پرتو های گاما پرداخته و به‌طور مستقیم اقدام به برآورد پارامترها مدل پیشنهادی کردند. نکته‌ای که در آن مقاله وجود دارد موضوع انتخاب متغیر و تأثیر تغییر توزیع پاسخ بررسی نشده است.

بعضی مواقع داده‌هایی وجود دارند که همگن و یک‌دست نیستند، در چنین شرایطی مدل‌های آمیزه‌ای متناهی به‌عنوان ابزاری انعطاف‌پذیر برای مدل‌بندی این‌گونه داده‌ها استفاده می‌شوند. ارمز و اسکندری (۲۰۱۶) بحث انتخاب متغیر با استفاده از شیوه‌های تاوانیده را در ترکیبی از مدل تعمیم‌یافتهٔ نیم‌پارامتری ارایه شده به‌وسیلهٔ لی و لیانگ (۲۰۰۸) و آمیزه‌ای متناهی از مدل‌های رگرسیونی خلیلی و چن (۲۰۰۷) مطرح کردند که تا آن زمان، شیوه‌های تاوانیده ارائه‌شده در مورد ترکیب دو مدل مطرح‌شده به‌طور هم‌زمان بررسی نشده بودند. چو و فریزلویز (۲۰۱۲) نیز موضوع انتخاب متغیر را برررسی کردند. در واقع آن‌ها بحث انتخاب متغیر در خانواده توزیع‌های نمایی را در مدل معرفی‌شده، به‌طور کلی بررسی کردند. هم‌چنین آمیزه‌ای متناهی از مدل‌های نیم‌پارامتری تعمیم‌یافته را با استفاده از برآورد تاوانیده مطرح کردند، در واقع مدل ربط را در حالت نیم‌پارامتری بسط دادند و تابع ناپارامتری را چندبعدی در نظر گرفتند.

آمیزه‌ای متناهی مدل‌های خطی تعمیم یافته با پاسخ توزیع دوجمله‌ای منفی

مدل آماری بر مبنای آمیزهٔ متناهی از توزیع‌های پواسون یک شیوه شناخته‌شده برای مدل‌بندی داده‌های شمارشی است. اما به‌دلیل ویژگی هم‌پراکنشی (برابر بودن میانگین و واریانس) در این توزیع، استفاده از آن در مسائل کاربردی به‌دلیل رخ دادن بیش‌پراکندگی محدود می‌شود. در نتیجه به‌دلیل وجود ساختار بیش‌پراکندگی، استفاده از روش‌های جای‌گزین بسیاری پیشنهاد شده است که منجر به توسعهٔ روش‌های آماری برای مدل‌بندی داده‌های شمارشی شده است. در این ساختار به‌عنوان جای‌گزین آمیزهٔ متناهی از توزیع‌های دوجمله‌ای متناهی پیشنهاد می‌شود. رگرسیون دوجمله‌ای منفی یک انتخاب مناسب برای مدل‌بندی رابطه بین متغیرهای توضیحی و یک متغیر وابسته شمارشی است. در واقع توزیع دوجمله‌ای منفی برای داده‌های شمارشی که بیش‌پراکنده هستند، قابل استفاده هستند. در چنین شرایطی می‌توان حالت تعمیم‌یافته رگرسیون پواسونی را در نظر گرفت، چون میانگین توزیع دوجمله‌ای منفی دارای ساختار یکسانی در مقایسه با رگرسیون پواسونی است, علاوه بر این یک پارامتر اضافه به‌منظور مدل کردن حالت بیش‌پراکندگی دارد. با ید توجه کرد شیوه‌ای علمی در مقابله با تغییرپذیری بسیار زیاد توزیع پواسون، استفاده از مدل آمیزه‌ای پیوسته پواسون است. در این شرایط و در حالت غیر رگرسیونی، میانگین پواسون(μ_j) به‌عنوان یک متغیر پنهان از یک توزیع به‌صورت $\mathrm{H}(\mu_j)$ در نظر گرفته می‌شود. بنابراین چگالی Y_j بدین‌صورت مدل‌بندی می‌شود:

$$f(y_j) = \int_0^\infty \{e^{-\mu}{\mu_j}^y / y_j!\} I_A(y_j) dH(\mu) \qquad (۵)$$

یک انتخاب رایج برای $H(\mu)$ توزیع گاما است که به‌صورت $\Gamma(\alpha, \beta)$ نمایش داده می‌شود. این انتخاب منجر می‌شود که رابطهٔ (۵) به‌صورت (۶) تعریف شود:

$$f(y_j; \alpha, \beta) = \binom{y_j+\alpha-1}{y_j} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^{y_j} I_A(y_j) \qquad (۶)$$

رابطهٔ (۶) دارای توزیع دوجمله‌ای منفی $NB(\alpha, \frac{\beta}{\beta+1})$ ، است.

انتخاب متغیر در آمیزه‌ای متناهی از مدل‌های نیم‌پارامتری دوجمله‌ای منفی

فرض کنید Y متغیر پاسخ دلخواه و (X, U, Z) برداری از متغیرهای کمکی تأثیرگذار بر متغیر پاسخ باشند. در چنین حالتی آمیزه‌ای متناهی از از یک مدل آمیزهٔ متناهی با پاسخ دوجمله‌ای منفی بدین‌صورت پیشنهاد می‌شود:

$$f(y_i; x_i, u_i, z_i, \Psi) = \sum_{k=1}^{K} \pi_k \, NB(\mu_{ik}, \phi_k) \quad (۷)$$

$$NB(\mu_{ik}, \phi_k) = \left[\frac{\Gamma(y_i + \phi_k)}{\Gamma(y_i + 1)\,\Gamma(\phi_k)}\left(\frac{\mu_{ik}}{\mu_{ik} + \phi_k}\right)^{y_i}\left(\frac{\phi_k}{\mu_{ik} + \phi_k}\right)^{\phi_k}\right] \quad i = 1, \dots, n; k = 1, \dots, K$$

در رابطهٔ (۷) بردار پارامتری Ψ به‌صورت $\Psi = (\beta_1, \beta_2, \dots, \beta_k, \alpha_1, \dots, \alpha_k, \phi)$ نمایش داده می‌شود که شامل $\beta_k = (\beta_{k1}, \dots, \beta_{kp})^T$، $\alpha_k = (\alpha_{k1}, \dots, \alpha_{kp})^T$ و $\pi = (\pi_1, \dots, \pi_{k-1})^T$ است، به‌طوری‌که $j = 1, \dots, P$ نمایش‌دهندهٔ ابعاد متغیرهای کمکی است. برای بردار π خاصیت $\pi_k > 0$ و $\sum_{k=1}^{K} \pi_k = 1$ برقرار است. در رابطهٔ (۷) μ_{ik} نرخ میانگین توزیع دوجمله‌ای منفی به‌ازای مشاهده iام و مؤلفهٔ kام است که با در رابطهٔ (۸) صدق می‌کند:

$$\mu_{ik}(x_i, u_i, z_i) = \exp(x_i^T \alpha_k(u_i) + z_i^T \beta_k) \quad (۸)$$

مسئلهٔ انتخاب متغیر را با رویکرد درستنمایی تاوانیده و با در نظر گرفتن آمیزه متناهی توزیع دوجمله‌ای منفی برای متغیر پاسخ، در سه مرحلهٔ اصلی و بر مبنای استفاده از الگوریتم EM بیان می‌کنیم.

مرحلهٔ اول: محاسبهٔ برآورد موضعی ضرایب ناپارامتری

لگاریتم تابع درست‌نمایی به شرط پارامتر Ψ برمبنای آمیزهٔ متناهی از توزیع‌های دوجمله‌ای منفی بدین‌صورت تعریف می‌شود:

$$\ell_n(\Psi) = \sum_{i=1}^{n} log\,\{\sum_{k=1}^{K} \pi_k \, f(y_i; \mu_{ik}(x_i, u_i, z_i), \phi_k)\} \quad (۹)$$

در نتیجه برای برآورد ضرایب مجهول باید برمبنای داده‌های کامل که در حضور متغیرهای نشان‌گر فرضی v_{ik} است، لگاریتم تابع درست‌نمایی کامل را بدین‌صورت تعریف کرد:

$$\ell_n^c(\Psi) = \sum_{i=1}^{n}\sum_{k=1}^{K} v_{ik}\{log\,\pi_k + log\{f(y_i; \mu_{ik}(x_i, u_i, z_i), \phi_k)\} \quad (۱۰)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} v_{ik}\{log\,\pi_k + log\left(\frac{\Gamma(y_i + \phi_k)}{\Gamma(y_i + 1)\,\Gamma(\phi_k)}\right) - (\phi_k + y_i)\,log(\mu_{ik} + \phi_k) + y_i\,log(\mu_{ik}) + \phi_k log\,(\phi_k)\}$$

به‌دلیل کارکردن با یک مدل آمیزه‌ای متناهی از توزیع‌های دوجمله‌ای منفی و عدم اطلاع در مورد تابع ناپارامتری $\alpha(.)$، برای برآورد بخش ناپارامتری از روش ارائه شده به‌وسیلهٔ لی و لیانگ (۲۰۰۸)، با استفاده از تقریب خطی مرتبهٔ اول تیلور، $\alpha_{kj}(v)$ به‌ازای v و در همسایگی متغیری مانند u به‌صورت (۱۱) ارایه می‌شود:

$$\alpha_{kj}(v) \approx \alpha_{kj}(u) + \alpha'_{kj}(u)(v - u) \equiv a_{kj} + b_{kj}(v - u) \quad (۱۱)$$

$$k = 1,2, \dots, K \quad j = 1,2, \dots, P$$

در رابطهٔ (۱۱) تابع $\alpha(.)$ به‌ازای متغیر u و مؤلفهٔ kام و بعد jام، با متغیر a_{kj} و مشتق مرتبهٔ اول آن با متغیر b_{kj} به‌منظور ساده‌سازی نمایش ارایه می‌شود.

a_{kj} و b_{kj} ضرایب پارامتری بسط تیلور هستند. بسط تیلور در رابطهٔ (۱۱) را تنها تا مرتبهٔ دوم در نظر گرفته‌ایم و این در حالی است که هر چه تعداد مرتبه افزایش یابد به‌دنبال آن دقت نیز افزایش خواهد یافت. به‌منظور هموارسازی بخش ناپارامتری از بسط سری تیلور استفاده می‌کنیم تا بتوانیم ضرایب را تقریب بزنیم.

بنابراین تابع درست‌نمایی برمبنای مؤلفهٔ ناپارامتری و تابع هسته $k_h(u_i-u)=\frac{1}{h}\,k(\frac{u_i-u}{h})$ ، برای برآورد موضعی و تقریبی a ، b و β به‌صورت (۱۲) تعریف می‌شود:

$$\ell_n=\sum_{i=1}^{n} log \sum_{k=1}^{K}\{\pi_k\left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\tilde{\mu}_{ik}}{\tilde{\mu}_{ik}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\tilde{\mu}_{ik}+\phi_k}\right)^{\phi_k}\right] k_h(u_i-u)\} \qquad (۱۲)$$

که در آن $\tilde{\mu}_{ik}(u_i,x_i,z_i)$ برمبنای a و b به‌صورت (۱۳) تعریف می‌شود:

$$\tilde{\mu}_{ik}(u_i,x_i,z_i)=\exp\left(x_i^T a_k+x_i^T b_k(u_i-u)+z_i^T\beta_k\right) \qquad (۱۳)$$

$$i=1,2,\dots,n \quad k=1,2,\dots,K$$

برای انجام عمل بیشینه‌سازی و محاسبهٔ مقدار بهینه برمبنای رابطهٔ (۱۲)، لگاریتم تابع درست‌نمایی کامل به‌صورت (۱۴) تعریف می‌شود:

$$\ell_n^c(\Psi)=\sum_{i=1}^{n}\sum_{k=1}^{K} v_{ik}\{log\,\pi_k+log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right)-(\phi_k+y_i)\,log(\tilde{\mu}_{ik}+\phi_k)+ y_i\,log(\tilde{\mu}_{ik})+\phi_k log\,(\phi_k)+log\;k_h(u_i-u)\} \qquad (۱۴)$$

در تابع هسته‌ی$k_h(u_i-u)$ ، h پهنای نوار است. پس از مشخص کردن $\tilde{\mu}_{ik}$ ، مقادیر بهینهٔ رابطهٔ (۱۴) را با استفاده از الگوریتم EM برآورد پارامترها و ضرایب بخش ناپارامتری را در ادامه محاسبه می‌کنیم.

گام E:

در این گام امید ریاضی شرطی $\ell_n^c(\Psi)$ به‌شرط v_{ik} و برمبنای مشاهدات (u_i,x_i,z_i,y_i) به‌صورت (۱۵) تعریف می‌شود:

(۱۵)

$$Q(\Psi;\Psi^{(m)})=\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}^{(m)}[log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right)-(\phi_k+y_i)\,log(\tilde{\mu}_{ik}+\phi_k)+y_i\,log(\tilde{\mu}_{ik})+\phi_k log\,(\phi_k)]+\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}^{(m)}\,log\pi_k\;+\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}^{(m)}\,log\;k_h(u_i-u)$$

به‌طوری‌که $\omega_{ik}^{(m)}$ ها امید ریاضی شرطیv_{ik} ها به شرط مشاهدات هستند و به‌عنوان مقادیر وزنی به‌صورت (۱۶) قابل دسترس هستند:

$$\omega_{ik}^{(m)}=\frac{\pi_k^{(m)}\left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\tilde{\mu}_{ik}{}^{(m)}}{\tilde{\mu}_{ik}{}^{(m)}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\tilde{\mu}_{ik}{}^{(m)}+\phi_k}\right)^{\phi_k}\right]}{\sum_{k=1}^{K}\pi_k^{(m)}\left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\tilde{\mu}_{ik}{}^{(m)}}{\tilde{\mu}_{ik}{}^{(m)}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\tilde{\mu}_{ik}{}^{(m)}+\phi_k}\right)^{\phi_k}\right]} \qquad (۱۶)$$

گام M:

در گام (m+1) امین مرحله از تکرار، $Q(\Psi;\Psi^{(m)})$ نسبت به اجزای بردار پارامتری Ψ بیشینه می‌شود. به هنگام استفاده از الگوریتم EM، بیشینه کردن $Q(\Psi;\Psi^{(m)})$ برمبنای نسبت‌های آمیزه‌ای به‌همراه دیگر پارامترها دارای پیچیدگی محاسباتی است. از این رو، به‌همین منظور لازم است در هر مرحله برمبنای وزن $\omega_{ik}^{(m)}$ ، نسبت‌های آمیزه‌ای $\pi_k^{(m+1)}$ به‌صورت (۱۷) محاسبه شود:

$$\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}\omega_{ik}^{(m)} \quad k = 1,2,\dots,K \qquad (۱۷)$$

به شرط مقدار $\pi_k^{(m+1)}$، می‌توان عبارت $Q(\Psi;\Psi^{(m)})$ را نسبت به a ، b و β ، بیشینه کرد. با توجه به رابطهٔ (۱۷)، a و b ضرایب بخش ناپارامتری هستند که با به‌کارگیری الگوریتم EM به‌دنبال برآورد تقریبی نتایج هستیم بنابراین به‌منظور تعیین برآورد ضرایب مربوط به بخش پارامتری و ناپارامتری، لازم است به حل معادله‌های (۱۸) و (۱۹) و (۲۰) بپردازیم:

$$\sum_{i=1}^{n}\omega_{ik}^{(m)}\frac{\partial}{\partial\beta_{kj}}\{log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right)-(\phi_k+y_i)\,log(\tilde{\mu}_{ik}+\phi_k)+y_i\,log(\tilde{\mu}_{ik})+$$
$$\phi_k log\,(\phi_k)+log\;k_h(u_i-u)\}=0 \qquad (۱۸)$$

$$\sum_{i=1}^{n}\omega_{ik}^{(m)}\frac{\partial}{\partial a_{kj}}\{log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right)-(\phi_k+y_i)\,log(\tilde{\mu}_{ik}+\phi_k)+y_i\,log(\tilde{\mu}_{ik})+$$
$$\phi klog(\phi k)+log khui-u\}=0 \qquad (۱۹)$$

$$\sum_{i=1}^{n}\omega_{ik}^{(m)}\frac{\partial}{\partial b_{kj}}\{\,log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right)-(\phi_k+y_i)\,log(\tilde{\mu}_{ik}+\phi_k)+y_i\,log(\tilde{\mu}_{ik})+$$
$$\phi_k log\,(\phi_k)+log\;k_h(u_i-u)\}=0 \qquad (۲۰)$$

چنان‌که از معادلات مذکور مشهود است، یک دستگاه پیچیده معادلاتی از مجهول‌ها ایجاد شده است، به‌گونه‌ای‌که حل آن به‌طور دستی امکان‌پذیر نیست و به‌طور صریح نمی‌توان فرم بسته‌ای برای آن‌ها به‌دست آورد. در ادامه به محاسبهٔ برآورد ها می‌پردازیم

مرحلهٔ دوم: محاسبه برآورد ضرایب $\boldsymbol{\beta}$ تاوانیده

در ابتدا به‌منظور برآورد ضرایب β تاوانیده، تابع درست‌نمایی تاوانیده (۲۱) تعریف می‌شود:

$$\ell(\beta)=\sum_{i=1}^{n}log\sum_{k=1}^{K}\{\pi_k\left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\mu_{ik}^{*}}{\mu_{ik}^{*}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\mu_{ik}^{*}+\phi_k}\right)^{\phi_k}\right]\}-p_n(\Psi) \qquad (۲۱)$$

که در آن μ_{ik}^{*} بدین‌صورت است:

$$\mu_{ik}^{*}(u_i,x_i,z_i)=exp(x_i^T\tilde{\alpha}_k(u_i)+z_i^T\beta_k) \qquad k=1,2,\dots,K \qquad (۲۲)$$

تفاوت μ_{ik}^{*} با $\tilde{\mu}_{ik}$ در این است که در رابطهٔ (۲۲) به‌جای تابع مجهول ضرایب ناپارامتری مرحلهٔ قبل، از برآورد بهینه حاصل از مرحلهٔ اول، استفاده می‌شود تا برآوردهای دقیق‌تری برای ضرایب پارامتری تاوانیده محاسبه شود. با بیشینه‌کردن تابع درست‌نمایی تاوانیده $\ell(\beta)$ نسبت به β، برآورد تاوانیده ضرایب بخش پارامتری به‌دست می‌آیند. در رابطهٔ (۲۲) به‌جای $p_{nk}(\beta)$ در $p_n(\Psi)$ از تقریب درجهٔ دوم موضعی آن در نزدیکی نقطه β به‌صورت رابطهٔ (۲۳) استفاده می‌کنیم.

$$p_n(\Psi)=\sum_{k=1}^{K}\pi_k\{\sum_{j=1}^{P}p_{nk}(\beta_{kj})\} \qquad (۲۳)$$

به‌طوری‌که داریم:

$$p_{nk}(\beta) \approx p_{nk}(\beta_0) + \frac{p'_{nk}(\beta)}{2\beta_0}(\beta^2 - \beta_0^2) \qquad (۲۴)$$

دلیل استفاده از تقریب رابطهٔ (۲۴) است که چون تابع تاوان $p_{nk}(\beta)$ در $\beta = 0$ مشتق‌پذیر نیست، با تقریب موضعی درجهٔ دوم تابع $p_{nk}(\beta)$ در نزدیکی نقطه صفر، مشکل عدم مشتق‌پذیری برطرف می‌شود. از این‌رو، به‌منظور محاسبه برآوردهای تاوانیده، تابع تاوان $p_n(\Psi)$ در مرحلهٔ m+1 ام را با مقدار تقریبی آن به‌صورت (۲۵) جای‌گزین می‌کنیم:

$$\tilde{p}_n(\Psi, \Psi^{(m)}) = \sum_{k=1}^{K} \pi_k \{\sum_{j=1}^{P} p_{nk}\left(\beta_{jk}^{(m)}\right) + \frac{p'_{nk}\left(\beta_{jk}^{(m)}\right)}{2\beta_{jk}^{(m)}}(\beta_{jk}^2 - \beta_{jk}^{(m)2})\} \qquad (۲۵)$$

گام E:

در این گام نیز مشابه گام محاسبهٔ امیدریاضی در مرحلهٔ اول، امیدریاضی شرطی لگاریتم تابع درست‌نمایی کامل تاوانیده $\tilde{\ell}_c(\Psi)$ به‌صورت $\tilde{\ell}_c(\Psi) = \ell_c(\Psi) - \tilde{p}_n(\Psi)$ را به‌شرط متغیرهای نشان‌گر v_{ik}ها و مشاهدات (u_i, x_i, z_i, y_i) محاسبه می‌کنیم، و داریم

$$Q(\Psi; \Psi^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{ik}^{(m)}[log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right) - (\phi_k + y_i)\,log(\mu_{ik}^* + \phi_k) +$$
$$y_i\,log(\mu_{ik}^*) + \phi_k log\,(\phi_k)] + \sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{ik}^{(m)}\,log\pi_k - \tilde{p}_n(\Psi) \qquad (۲۶)$$

که در آن $\omega_{ik}^{(m)}$ به‌صورت (۲۷) نمایش داده می‌شود:

$$\omega_{ik}^{(m)} = \frac{\pi_k^{(m)}\left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\mu_{ik}^{*\,(m)}}{\mu_{ik}^{*\,(m)}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\mu_{ik}^{*\,(m)}+\phi_k}\right)^{\phi_k}\right]}{\sum_{k=1}^{K}\pi_k^{(m)}\left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\mu_{ik}^{*\,(m)}}{\mu_{ik}^{*\,(m)}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\mu_{ik}^{*\,(m)}+\phi_k}\right)^{\phi_k}\right]} \qquad (۲۷)$$

اکنون به‌جای $\tilde{p}_n(\Psi)$ برمبنای تابع‌های تاوان LASSO ، و یا SCAD ، از تقریب درجهٔ دوم موضعی آن‌ها می‌توان استفاده کرد.

گام M:

در گام (m+1) امین مرحله از تکرار، $Q(\Psi; \Psi^{(m)})$ تعریف‌شده در گام E را با وجود برآورد موضعی بخش ناپارامتری، نسبت به پارامترهای مجهول آن بیشینه می‌کنیم. همانند مرحلهٔ اول، در این مرحله نیز پس از به روزآوری نسبت‌های آمیزه‌ای، با ثابت فرض کردن π_k در $Q(\Psi; \Psi^{(m)})$ ، آن را نسبت به β بیشینه می‌کنیم و داریم:

$$\sum_{i=1}^{n} \omega_{ik}^{(m)} \frac{\partial}{\partial\beta_{kj}}\{log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right) - (\phi_k + y_i)\,log(\mu_{ik}^* + \phi_k) + y_i\,log(\mu_{ik}^*) +$$
$$\phi_k log\,(\phi_k)\} - \pi_k\{\frac{\partial}{\partial\beta_{kj}}\tilde{p}_n(\beta_{kj})\} = 0 \qquad (۲۸)$$

در واقع در گام M نیز به‌جای $\tilde{p}_n(\beta_{kj})$ به‌طور جداگانه می‌توان با توجه به نوع تابع تاوان به‌کار رفته در گام E از توابع تاوان $\tilde{p}_n^L(\Psi; \Psi^{(m)})$ و یا $\tilde{p}_n^S(\Psi; \Psi^{(m)})$ به‌ازای مقادیر $k = 1,2, \dots, K$ و $j = 1,2, \dots, P$ استفاده کرد.

باید توجه کرد مقدار بهینه زمانی به‌دست می‌آید که پس از تکرارهای مراحل مختلف به‌ازای گام‌های E و M ، اختلاف نرم اقلیدسی به‌ازای مقادیر برآورد ضرایب پارامتری برای دو مرحله متوالی به‌صورت $\left\|\beta_{11}^{(m+1)} - \beta_{11}^{(m)}\right\|$ کم‌تر از

مقدار کوچک دلخواهی همانند δ شود. در چنین مرحله‌ای برآورد ضرایب پارامتری تاوانیده $\hat{\beta}$ به‌دست می‌آید و می‌توان هم‌گرایی را تأیید کرد.

مرحلهٔ سوم: محاسبهٔ برآورد دقیق ضرایب ناپارامتری

در این مرحله از برآوردهای تاوانیده ضرایب پارامتری $\hat{\beta}$ حاصل از مرحله می‌توان استفاده کرد و آن‌ها را به‌جای برآوردهای موضعی غیرتاوانیده مرحلهٔ اول قرار می‌دهیم. در واقع در این مرحله $\tilde{\mu}_{ik}^*$ ، به‌جای $\tilde{\mu}_{ik}$ در مرحلهٔ اول جای‌گزین می‌شود و به‌صورت (۲۹) تعریف می‌شود:

$$\tilde{\mu}_{ik}^* = exp\left(x_i^T a_k + x_i^T b_k(u_i - u) + z_i^T \hat{\beta}_k\right) \quad (۲۹)$$

در واقع در تعریف $\tilde{\mu}_{ik}^*$، ضرایب بخش ناپارامتری a و b مجهول در نظر گرفته می‌شوند و برآورد پارامتری ضرایب رگرسیونی مرحلهٔ دوم جای‌گذاری می‌شود. در این مرحله هدف اصلی محاسبه برآورد دقیق a و b به‌جای برآوردهای موضعی $\tilde{a}$ و $\tilde{b}$ مرحلهٔ اول در حضور برآورد های تاوانیده ضرایب رگرسیونی است. تابع درست‌نمایی در این مرحله براساس $\tilde{\mu}_{ik}^*$ بدین‌صورت است:

$$\sum_{i=1}^{n} log \sum_{k=1}^{K} \{\pi_k \left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\tilde{\mu}_{ik}^*}{\tilde{\mu}_{ik}^*+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\tilde{\mu}_{ik}^*+\phi_k}\right)^{\phi_k}\right] \quad (۳۰)$$

$$k_h(u_i - u)\}$$

در نتیجه تابع درست‌نمایی کامل در حضور متغیرهای ν_{ik} بدین‌صورت تعریف می‌شوند:

$$\ell_n^c(\Psi) = \sum_{i=1}^{n}\sum_{k=1}^{K} \nu_{ik}\{log\,\pi_k + [log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right) - (\phi_k + y_i)\,log(\tilde{\mu}_{ik}^* + \phi_k) +$$

$$yilog\mu ik* + \phi klog(\phi k)] + log khui - u\} \quad (۳۱)$$

گام E

امیدریاضی شرطی $\ell_n^c(\Psi)$ به‌شرط متغیرهای نشان‌گر نامشهود ν_{ik} و مشاهدات (u_i, x_i, z_i, y_i) بدین‌صورت تعریف می‌شود:

(۳۲)

$$Q\left(\Psi; \Psi^{(m)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{ik}^{(m)} [log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right) - (\phi_k + y_i)\,log(\tilde{\mu}_{ik}^* + \phi_k)$$

$$+ y_i\,log(\tilde{\mu}_{ik}^*) + \phi_k \log(\phi_k)]$$

$$+ \sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{ik}^{(m)}\,log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{ik}^{(m)} \log k_h(u_i - u)$$

که در آن $\omega_{ik}^{(m)}$ها به‌صورت (۳۳) محاسبه می‌شوند.

$$\omega_{ik}^{(m)} = \frac{\pi_k^{(m)}\left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\tilde{\mu}_{ik}^{*(m)}}{\tilde{\mu}_{ik}^{*(m)}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\tilde{\mu}_{ik}^{*(m)}+\phi_k}\right)^{\phi_k}\right]}{\sum_{k=1}^{K}\pi_k^{(m)}\left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\left(\frac{\tilde{\mu}_{ik}^{*(m)}}{\tilde{\mu}_{ik}^{*(m)}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\tilde{\mu}_{ik}^{*(m)}+\phi_k}\right)^{\phi_k}\right]} \quad (۳۳)$$

گام M:

در گام (m+1) امین مرحله از تکرار، $Q(\Psi;\Psi^{(m)})$ را نسبت به ضرایب مجهول بخش ناپارامتری یعنی a و b بیشینه می‌کنیم. در واقع این مرحله بعد از به روزآوری نسبت‌های آمیزه‌ای انجام می‌پذیرد. بنابراین با ثابت در نظر گرفتن π_k و با حل معادلات (۳۴) و (۳۵) برآوردهای مورد نظر به‌دست می‌آیند:

$$\sum_{i=1}^{n}\omega_{ik}^{(m)}\frac{\partial}{\partial a_{kj}}\{log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right)-(\phi_k+y_i)\,log\left(\tilde{\mu}_{ik}^{*(m)}+\phi_k\right) + y_i\,log\left(\tilde{\mu}_{ik}^{*(m)}\right)+\phi_k log\,(\phi_k)+log\;k_h(u_i-u)\} = 0 \quad (۳۴)$$

$$\sum_{i=1}^{n}\omega_{ik}^{(m)}\frac{\partial}{\partial b_{kj}}\{log\left(\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\,\Gamma(\phi_k)}\right)-(\phi_k+y_i)\,log\left(\tilde{\mu}_{ik}^{*(m)}+\phi_k\right) + y_i\,log\left(\tilde{\mu}_{ik}^{*(m)}\right)+\phi_k log\,(\phi_k)+log\;k_h(u_i-u)\} = 0 \quad (۳۵)$$

بعد از حل معادلهٔ (۳۵) در گام M از مرحلهٔ دوم، به‌منظور برآورد دقیق‌تر ضرایب مجهول بخش ناپارامتری در مقابل برآوردهای موضعی ضرایب ناپارامتری در مرحلهٔ اول، به حل معادلات (۳۴) و(۳۵) می‌پردازیم. در این‌جا نیز همانند مرحلهٔ اول با یک دستگاه پیچیده معادلاتی مواجه هستیم که نمی‌توان به‌طور صریح، فرم بسته‌ای برای آن به‌دست آورد. در نتیجه جواب بهینه دستگاه معادلات را با تعداد تکرار لازم به‌دست می‌آوریم.

زمانی که الگوریتم EM پس از طی مراحل E و M به‌هم‌گرایی برسد، برآوردهای اصلی $\hat{a}$ و $\hat{b}$ به‌دست می‌آیند. بنابراین در مرحلهٔ سوم برآوردهای دقیق ضرایب ناپارامتری و از مرحلهٔ دوم برآوردهای تاوانیده ضرایب پارامتری به‌صورت $\{\hat{a},\hat{b},\hat{\beta}\}$ حاصل می‌شوند. انجام محاسبات به‌طور کلی در این پنج گام انجام می‌پذیرد:

الف) در نظر گرفتن یک مقدار اولیه برای β تحت عنوان β^0،

ب) محاسبه برآورد موضعی تابع ناپارامتری به صورت $\tilde{\alpha}(u)=\tilde{a}$،

ج) محاسبه برآورد موضعی تابع ناپارامتری به صورت $\tilde{\alpha}(u)=\tilde{a}$ ،

د) محاسبه برآورد ضرایب پارامتری تاوانیده β تحت عنوان $\hat{\beta}$، در حضور تابع ناپارامتری $\tilde{\alpha}(u)$ و

ه) محاسبه برآورد دقیق تابع ناپارامتری بر مبنای ضرایب a و b بسط تیلور در حضور ضرایب تاوانیده $\hat{\beta}$، تحت عنوان $\hat{a}$ و $\hat{b}$.

در هر یک از گام‌های محاسبات، نتایج نهایی پس از تکرار مراحل E و M و هم‌گرایی الگوریتم EM در نظر گرفته می‌شوند.

بررسی شبیه‌سازی

در این قسمت به‌منظور مقایسهٔ نتایج حاصل از استفاده از توزیع دوجمله‌ای منفی به‌جای توزیع پواسون در آمیزه‌ای متناهی از مدل‌ها، با استفاده از انجام یک برنامه شبیه‌سازی اقدام به برآورد پارامترها می‌کنیم. به‌این منظور این مراحل را انجام می‌دهیم:

۱. ابتدا به‌عنوان مقدار اولیه، بردار پارامترها را به‌صورت $\beta^0 = \begin{pmatrix} 1 & .7 \\ .5 & -.8 \end{pmatrix}$ در نظر گرفته و الگوریتم شبیه‌سازی را ۱۰۰ بار تکرار می‌کنیم. از آن‌جاکه براساس هر مقدار اولیه می‌توان هر دو مدل را بررسی کرد، از این‌رو، تأثیری در نتایج برای مقایسه دو مدل نخواهد داشت.

۲. در این مرحله دو توزیع پواسون را به‌صورت ترکیب خطی نوشته، به طوری که وزن هر دو توزیع پواسون با استفاده از الگوریتم EM به‌صورت (۰/۴۵۰۱۴۶و ۰/۵۴۹۸۵۴) خواهد بود. همین موضوع در خصوص دو توزیع دو جمله‌ای منفی به‌ترتیب به‌صورت (۰/۴۹۲۱۹۶ و ۰/۵۰۷۸۰۴) است.

۳. در این مرحله لگاریتم تابع درستنمایی برای مشاهده‌ها براساس آمیزه‌ای از دو توزیع پواسون و آمیزه‌ای از دو توزیع دوجمله‌ای منفی را به‌دست می‌آوریم. نتایج ارایه شده در جدول ۱ براساس لگاریتم تابع درستنمایی می‌توان نتیجه گرفت که آمیزه‌ای از دوتوزیع دوجمله‌ای منفی از درجه اعتبار بیش‌تری نسبت به آمیزه‌ای از دو توزیع پواسون برخوردار است مقدار ارایه شده اختلاف چشم‌گیری را اعلام می‌دارد.

جدول ۱. برآورد مقادیر لگاریتم تابع درستنمایی برای آمیزه‌ای از توزیع‌ها

برآورد لگاریتم تابع درستنمایی	نوع آمیزه‌ای از توزیع‌ها
-۳۰۷/ ۳۷۱۸	ترکیب دو توزیع پواسون
-۲۸۹/۹۰۹۴	ترکیب دو توزیع دوجمله‌ای منفی

با توجه به جدول ۱ درمی‌یابیم که مقدار آماره لگاریتم تابع درستنمایی در به‌کارگیری از توزیع دوجمله‌ای منفی نسبت به توزیع پواسون مقداری بزرگ‌تر دارد. این موضوع بیان می‌کند برازش مدل در استفاده از آمیزه‌ای متناهی از مدل‌های خطی تعمیم یافته بر مبنای توزیع دوجمله‌ای منفی نسبت به آمیزه متناهی از مدل‌های خطی تعمیم یافته بر مبنای توزیع پواسون بهتر انجام می‌پذیرد.

هم‌چنین جدول ۲ برآورد پارامترها و برآورد واریانس برآورد کننده‌ها در هر دو توزیع آمیزه‌ای ارایه داده است. آماره‌های والد ارایه شده برای برآورد کننده‌های تعیین شده در هر دو توزیع آمیزه‌ای هم‌چنان تایید فرض‌های مربوط را برای آمیزه‌ای متناهی دو توزیع دو جمله‌ای منفی بیان می‌کند.

جدول ۲. برآورد پارامترها و تعیین آماره آزمون براساس دو توزیع پواسون و دوجمله‌ای منفی

معنی‌داری	آمارهٔ والد	برآورد واریانس برآوردکننده	برآورد پارامترهای	نوع توزیع
فرض رد می‌شود	۱۴/۱۰۹	۱/۰۶۷	۳/۸۸	توزیع اول پواسون
فرض رد می‌شود	۴/۹۸	۰/۰۰۰۰۵	۰/۴۹۹	توزیع دوم پواسون
فرض مورد تایید است	۰/۸۷۴	۰/۳۵۶	-۰/۵۵۸	توزیع اول دوجمله‌ای منفی
فرض مورد تایید است	۴/۸۷	۰/۰۰۰۰۶	۰/۵۴۱	توزیع دوم دوجمله‌ای منفی

اما برای بررسی این‌که آیا تغییر توزیع‌ها برای آمیزه‌ای متناهی از ترکیب دو توزیع پواسون به آمیزه‌ای متناهی از ترکیب دو توزیع دوجمله‌ای منفی است یک کار مناسبی است از آمارهٔ آزمون میانگین وزنی تعمیم یافته توان دوم خطا استفاده می‌کنیم.که آن را با $WGMSE$ نشاده داده و بدین‌صورت تعریف می‌کنیم:

$$\mathrm{WGMSE} = \gamma\Lambda_1 + (1-\gamma)\Lambda_2$$

به‌طوری‌که در آن برای $i= 1,2$ داریم

$$\Lambda_i = \left(\widehat{\beta_{i0}}, \widehat{\beta_{i1}}\right)^T \begin{pmatrix} Var\left(\widehat{\beta_{i0}}\right) & COV\left(\widehat{\beta_{i0}}, \widehat{\beta_{i1}}\right) \\ COV\left(\widehat{\beta_{i0}}, \widehat{\beta_{i1}}\right) & Var\left(\widehat{\beta_{i1}}\right) \end{pmatrix} \left(\widehat{\beta_{i0}}, \widehat{\beta_{i1}}\right)$$

با توجه به این‌که مقدار وزن‌های برآورد شده در آمیزه‌ای متناهی از توزیع‌های پواسون عبارت است از:

$$\gamma^{pois} = (0/55\ ,0/45\ )$$

و مقادیر Λ_i برای توزیع‌ها پواسون عبارت است از

$$\Lambda^{pois} = (0/783,\quad 0/736)$$

از این‌رو، داریم:

$$WGMSE^{pois} = (0/55 * 0/783) + (0/45 * 0/736) = 0/762$$

اگر این موضوع را برای آمیزه‌ای متناهی از توزیع‌های دوجمله‌ای منفی در نظر بگیریم داریم:

$$\gamma^{nbinom} = (0/43\ ,0/57\ )$$

و $\Lambda^{nbinom} = (0/105,\ 0/068)$ است. در نتیجه می‌توان نوشت:

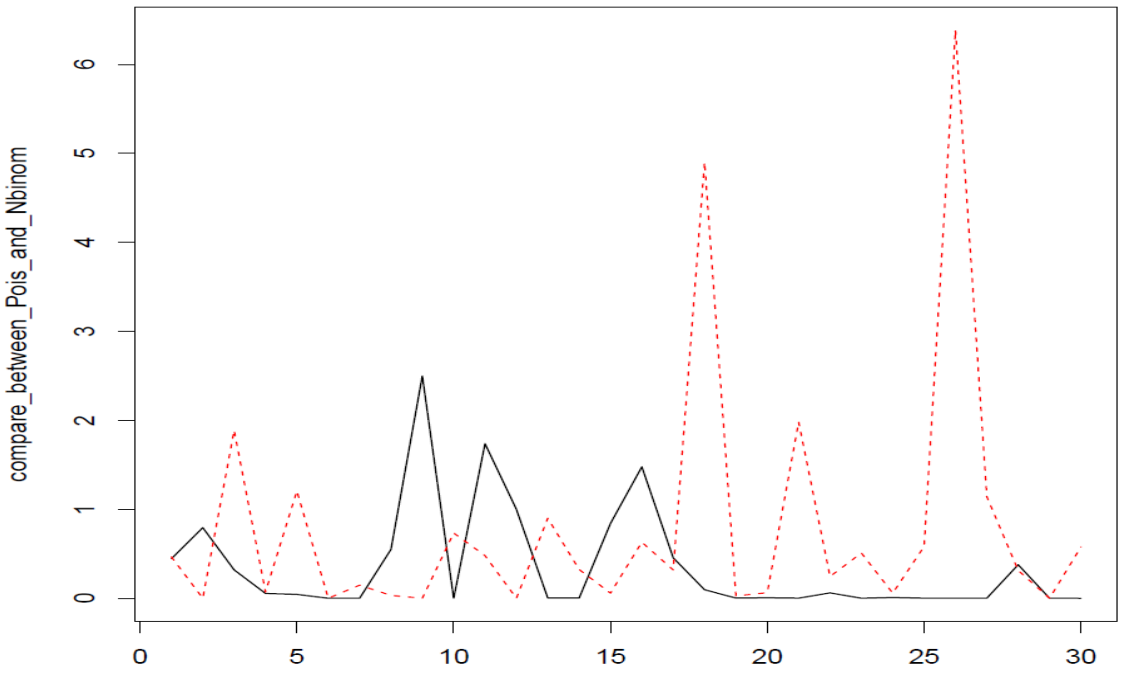
$$WGMSE^{nbinom} = (0/43 * 0/105) + (0/57 * 0/068) = 0/083$$

چنان‌که دیده می‌شود براساس هر دو توزیع دوجمله‌ای منفی که جای‌گزین دو توزیع پواسون شده است مقدار مقایسه‌ای آماره های آزمون آزمون کوچک‌تر شده است. در واقع می توان نوشت:

$$\text{کارایی آمیزه ای دوجمله ای منفی نسبت به آمیزه ای توزیع پواسون} = \frac{0/762}{0/083} = 9/18$$

نمودار ۱ نیز براساس ۳۰ بارتکرار شبیه‌سازی مدل پیشنهادی برتری انتخاب توزیع‌های دوجمله‌ای منفی را بر توزیع‌های پواسون نشان می‌دهد. در نمودار ۱، نمودار نقطه‌چین مربوط به مقادیر Λ در توزیع پواسون و نمودار پیوسته مربوط به توزیع پواسون بعد از ۳۰ بار تکرار است.

[DOR: 20.1001.1.25882546.1398.5.1.6.8]
[DOI: 10.29252/mmr.5.1.9]



نمودار ۱. مقایسهٔ بین مقادیر Δ بین دوتوزیع دوجمله‌ای منفی و توزیع پواسون

مثال کاربردی

یکی از کاربردهای مهم آمیزه‌ای متناهی از مدل‌های آماری استفاده از آن در مفهوم هم‌ترازسازی است. پژوهش در زمینهٔ هم‌ترازسازی آزمون‌ها، تاریخچهٔ طولانی دارد. بیش‌ترین کار قابل‌توجه مربوط به مک‌لاکلان و پیل (۲۰۰۰) مرور جامعی بر مدل‌های آمیزه‌ای متناهی داشته است. کلر و همکاران (۲۰۰۴) توانایی چهار روش هم‌ترازسازی (پرسش هم‌زمان، میانگین-سیگما، منحنی ویژگی آزمون و هم‌ترازسازی پارامتر پرسش مشترک ثابت) را به‌منظور بهبود تغییرات در توزیع توانایی آزمودنی با استفاده از داده‌های شبیه‌سازی مبنی براین‌که توانایی دارای توزیع نرمال استاندارد باشد، ارزیابی کردند. آن‌ها دریافتند که روش میانگین-سیگما بهترین روش است و روش هم‌ترازسازی پارامتر پرسش مشترک ثابت بدترین روش است. ون‌داویر و همکاران (۲۰۰۴) هم‌ترازسازی هسته را به‌طور کامل تعمیم دادند و روشی سیستماتیک برای بسیاری از طرح‌های هم‌ترازسازی معروف برای دست‌یابی به توابع هم‌ترازساز ارائه کردند. هو (۲۰۰۸) در یک بررسی شبیه‌سازی، ۱۰ تغییر حاصل از چهار روش هم‌ترازسازی (پرسش هم‌زمان، میانگین-سیگما، منحنی ویژگی آزمون و هم‌ترازسازی پارامتر پرسش مشترک ثابت) را در صورت وجود و نبود دور افتاده‌ها در مجموعهٔ پرسش‌های مشترک بررسی کرد. نتایج کار وی نشان داد که تبدیل‌های میانگین-سیگما و منحنی ویژگی آزمون از همه بهتر عمل می‌کند و پرسش‌های هم‌زمان و هم‌ترازسازی پارامتر پرسش مشترک ثابت دارای اثر متقابل پیچیده با گروه هم‌تراز و تعداد نقاط امتیاز دورافتاده‌ها است.

لی و بان (۲۰۱۰) چهار روش مختلف هم‌ترازسازی نظریه پرسش پاسخ (پرسش هم‌زمان، منحنی ویژگی آزمون، هابرا و تبدیل کارایی) را با هم مقایسه کردند و دریافتند که روش‌های پرسش جداگانه بهتر از روش‌های پرسش

هم‌زمان و تبدیل کارایی عمل می‌کنند. ون‌داویر (۲۰۱۱) یک روش مدل‌سازی آماری کلی را معرفی کرد که چارچوبی را برای بسیاری از روش‌های پیوند مقیاس فراهم می‌کند.

در این پژوهش جامعهٔ آماری بررسی شده مربوط به تعداد شرکت کنندگان در آزمون مقطع دکتری رشته آمار در دو سال متوالی ۹۲ و ۹۳ است که به‌صورت سراسری به‌وسیلهٔ سازمان سنجش آموزش کشور برگزار شده است. هریک از این آزمون‌ها شامل ۴۵ پرسش هستند. در آزمون مقطع دکتری رشته آمار سال ۹۲، ۱۰۸۳ نفر و در آزمون مقطع دکتری رشته آمار سال ۹۳، ۱۱۳۰ نفر شرکت کرده بودند و از اطلاعات همه این افراد در این پژوهش استفاده شده است. هم‌ترازی در این پژوهش مانند آمیزه‌ای متناهی از مدل آماری دوجمله‌ای منفی خواهد بود. با توجه به این‌که پاسخ‌ها در دو وضعیت درست و یا غلط در هر گزینه خواهد بود و داوطلبان باید آنقدر به پرسش‌ها پاسخ درست بدهند تا حد نصاب لازم برای پذیرش را داشته باشند از این‌رو، از لحاظ نظری پرسش‌های مورد نظر از ساختار توزیع دوجمله‌ای منفی تبعیت می‌کنند. در نتیجه از این پس به‌طور مترادف آن‌را به‌کار می‌بریم. در این پژوهش یک بار داده‌ها را براساس آمیزه متناهی از توزیع‌های پواسون و یک بار براساس آمیزه متناهی از توزیع‌های دوجمله‌ای منفی داده‌ها را تحلیلی می‌کنیم. برای استفاده از هم‌تراز ساز بودن است پرسش‌های مشترک لنگر را داشته باشیم. پرسش‌های لنگر دو جامعه در جدول ۴ آورده شده است.

جدول۴: تعداد پرسش‌های مشابه آزمون ۹۲ و آزمون ۹۳

آمار ۹۳	آمار ۹۲
۵	۵
۶	۷
۱۰	۸
۱۱	۱۱
۲۱	۱۳
۳۲	۳۰
۳۳	۳۱
۳۴	۳۳
۳۵	۳۵
۳۹	۳۸
۴۱	۴۲

لازم به ذکر است به‌منظور درک علت دقیق اطلاعات شرح داده شده باید به تحلیل گزینه‌های هر پرسش پرداخت. ممکن است گزینه‌های انحرافی به درستی انتخاب نشده باشند و یا برعکس گزینه‌های انحرافی به خوبی توانسته باشند کار خود را انجام دهند و نتایج قابل قبولی را ایجاد نمایند.

در این بخش ابتدا توصیفی مقدماتی از پاسخ نامه‌ها که دربرگیرندهٔ شاخص‌های آماری کلاسیک نیز می‌شود ارائه می‌شود، سپس ضرایب هم‌ترازسازی با استفاده از روش‌های گشتاوری و منحنی ویژگی به‌دست می‌آوریم و با استفاده از این ضرایب، مقادیر تیمارهای سال ۹۲ در مقیاس مقادیر تیمارهای سال ۹۳ قرار می‌گیرند.

تحلیل داده‌ها تنها براساس امتیاز کل انجام می‌شود بدین معنی که امتیاز یک‌سان بر اساس تعداد پاسخ‌های درست داوطلب توجه به دشواری و سادگی پرسش است، درنتیجه دو فرد که به تعداد یک‌سان به پرسش‌های مختلف از لحاظ دشواری پاسخ داده‌اند امتیاز یک‌سان داده می شود.

جدول ۵. آماره‌های توصیفی مربوط به تعداد کل پاسخ‌ها

	مقادیرسال ۹۲	مقادیر سال ۹۳
مینیمم	۰	۰
ماکسیمم	۳۲	۳۰
میانگین	۱۴/۱۳۸	۱۴/۵۳
میانه	۱۴	۱۴/۵
مد	۱۰	۱۱
انحراف معیار	۸/۷۸	۸/۸۶

آن‌چه که از جدول ۵ دریافت می‌شود این است که کم‌ترین امتیاز کل کسب شده به‌وسیلهٔ همهٔ داوطلبان در هر دو جامعه، صفر است. سایر آماره‌های مربوط به هر دو جامعه خیلی نزدیک به هم هستند. می‌توان نتیجه گرفت هر دو جامعه از لحاظ دشواری و سطح پرسش‌ها یک‌سان هستند.

آماره مربوط به سطح اندازه‌گیری آزمون است آماره پایایی است. پایایی آزمون در نظریه آزمون کلاسیک از طریق آلفای کرونباخ محاسبه شده است. نتایج پایایی برای هر دو جامعه براساس ضریب آلفای کرونباخ به‌ترتیب برابر با ۰/۷۲ برای آزمون سال ۹۲ و ۰/۷۰ برای آزمون سال ۹۳ است. یکی از شرایط هم‌ترازسازی این است که هر دو جامعه دارای پایایی مشابهی باشند. بدیهی است که این شرط برقرار است.

آماره‌های پرسش در نظریه آزمون کلاسیک عبارت از ضریب دشواری و ضریب قدرت تمیز است. بر اساس تعریف ضریب دشواری هر پرسش برابر است با نسبت داوطلبانی که به پرسش مورد نظر پاسخ درست داده‌اند. براساس تعریف ضریب قدرت تمیز نیز همان هم‌بستگی بین هر پرسش و مقدار امتیاز کل است. در جدول ۶ مقادیر مربوط به این آماره‌ها برای ۶ پرسش مشابه در دو آزمون سال ۹۲ و ۹۳ ارایه شده است. روشن است ضریب دشواری که به‌منظور تشخیص پرسش‌های دشوار از آسان به‌کار می‌رود هر چه بیش‌تر باشد به معنی ساده‌تر بودن پرسش است. در نتیجه براساس جدول ۶ مشاهده می‌شود که هیچ‌یک از پرسش‌های این دو آزمون ساده نیستند. هم‌چنین در آزمون سال ۹۲ بر خلاف آزمون سال ۹۳ ، تعداد پرسش‌های دشوار بیش‌تر از تعداد پرسش‌های با سطح دشواری مناسب است. از طرفی ضریب قدرت تمیز پرسش که برای تشخیص افراد قوی از ضعیف استفاده می‌شود تنها برای تعداد کمی از پرسش‌ها در هر دو جامعه، مقدار بیش‌تر از ۰/۴ دارد. با توجه به این اطلاعات بهتر است روی پرسش‌هایی که ضریب تمیز اندکی دارند و نیز پرسش‌هایی که ضریب دشواری خیلی زیاد و یا خیلی اندک دارند بررسی بیش‌تری انجام گیرد و در صورت لزوم حذف شوند تا بتوان به نتایج دقیق‌تری دست پیدا کرد. بررسی های انجام شده نشان می دهد ضریب هم‌بستگی پیرسون بین ضریب دشواری و ضریب قدرت تمیز برای سال‌های ۹۲ و ۹۳ به‌ترتیب ۰/۳۷۰ - و ۰/۳۳ است. هم‌بستگی مثبت برای سال ۹۳ بدین‌معنی است که در سایر پاسخ‌ها با افزایش (یا کاهش) ضریب دشواری، ضریب قدرت تمیز نیز افزایش (یا کاهش) می‌یابد.

بعد از ارایه نتایج توصیفی داده‌ها، اکنون می‌خواهیم مدل‌های آمیزه‌ای متناهی توزیع‌های پواسون و آمیزه‌ای متناهی توزیع‌های دوجمله‌ای منفی را به داده‌ها برازش دهیم. برای این منظور در ابتدا بعد از برازش هر دو مدل برآورد لگاریتم تابع درست‌نمایی هر دو مدل را به‌دست می‌آوریم. در جدول ۷ این مقادیر آورده شده است. نتایج ارایه

شده در جدول ۷ برتری استفاده از آمیزه‌ای متناهی از مدل‌های خطی تعمیم یافته بر مبنای توزیع دوجمله‌ای منفی نسبت به آمیزه‌ای متناهی از مدل‌های خطی تعمیم یافته بر مبنای توزیع پواسون نشان داده شده است.

جدول ۶. آماره‌های کلاسیک دشواری و قدرت تمیز پرسش‌های مشابه سال ۹۲ و ۹۳

پرسش	سال ۹۲		سال ۹۳	
	دشواری	قدرت تمیز	دشواری	قدرت تمیز
۱	۰/۳۰۹۳	۰/۲۴۲۳	۰/۲۲۳۹	۰/۲۳۹۳
۲	۰/۲۹۷۳	۰/۳۰۲۲	۰/۳۸۴۱	۰/۴۲۶۴
۳	۰/۰۶۴۶	۰/۱۳۲۱	۰/۲۴۳۴	۰/۳۳۷۴
۴	۰/۲۹۶۴	۰/۲۲۴۷	۰/۱۸۱۴	۰/۱۸۸۸
۵	۰/۰۶	۰/۱۳۹۶	۰/۱۸۲۳	۰/۱۷۱۱
۶	۰/۰۵۶۳	۰/۰۹۰۷	۰/۰۷۰۸	۰/۰۹۸

جدول ۷. برآورد مقادیر لگاریتم تابع درست‌نمایی برای آمیزه‌ای از توزیع‌ها

برآورد لگاریتم تابع درست‌نمایی	نوع آمیزه‌ای از توزیع‌ها
-۱۳۲/۷۹۶۲	ترکیب دو توزیع پواسون
-۴۸/۹۰۹۴	ترکیب دو توزیع دوجمله‌ای منفی

جدول ۸ برآورد پارامترهای مدل و برآورد واریانس برآورد کننده‌ها ارایه شده است. آماره‌های والد ارایه شده اعلام می دارد که فرض‌های عدم تأثیر مقادیر ثابت و عامل کمی بر متغیر پاسخ تأثیرگذار است.

جدول ۸. برآورد پارامترها و تعیین آماره آزمون براساس توزیع‌های پواسون و توزیع‌های دوجمله‌ای منفی

معنی داری	آماره والد	برآورد واریانس برآورد کننده	برآورد پارامتر	نوع توزیع
فرض مورد تایید است	۱/۲۰۴	۰/۰۰۱	۰/۰۳۸۱	توزیع اول پواسون
فرض مورد تایید است	۰/۶۹۵	۰/۰۰۱	۰/۰۲۲	توزیع دوم پواسون
فرض مورد تایید است	۰/۹۱۷	۰/۰۰۱	۰/۰۲۹	توزیع اول دوجمله‌ای منفی
فرض مورد تایید است	۲/۰۲	۰/۰۰۱	۰/۰۶۴	توزیع دوم دوجمله‌ای منفی

برای بررسی این‌که آیا تغییر توزیع آمیزه‌ای متناهی از ترکیب دو توزیع پواسون به آمیزه‌ای متناهی از ترکیب دو توزیع دوجمله‌ای منفی مناسب خواهد بود از آماره آزمون میانگین وزنی تعمیم یافته توان دوم خطا ($WGMSE$) استفاده می‌کنیم که عبارت است از:

$$\mathrm{WGMSE} = \gamma\Lambda_1 + (1-\gamma)\Lambda_2$$

به‌طوری‌که در آن برای $i= 1,2$ داریم:

$$\Lambda_i = \left(\widehat{\beta_{i0}}, \widehat{\beta_{i1}}\right)^T \begin{pmatrix} Var\left(\widehat{\beta_{i0}}\right) & COV\left(\widehat{\beta_{i0}}, \widehat{\beta_{i1}}\right) \\ COV\left(\widehat{\beta_{i0}}, \widehat{\beta_{i1}}\right) & Var\left(\widehat{\beta_{i1}}\right) \end{pmatrix} \left(\widehat{\beta_{i0}}, \widehat{\beta_{i1}}\right)$$

با توجه به این‌که مقدار وزن‌های برآورد شده در آمیزه‌ای متناهی از توزیع‌های پواسون عبارت است از:

$$\gamma^{pois} = (0/50\ ,0/50\ )$$

و مقاد یر Λ_i برای توزیع ها پواسون عبارت است از:

$$\Lambda^{pois} = (0.432, \quad 0.697)$$

از این‌رو، داریم:

$$WGMSE^{pois} == (0/50 * 0/432) + (0/50 * 0/697) = 0/564$$

اگر این موضوع را برای آمیزه‌ای متناهی از توزیع‌های دوجمله‌ای منفی در نظر بگیریم داریم::

$$\gamma^{nbinom} = (0/43\,, 0/57\ )$$

و $\Lambda^{nbinom} = (1/016,\ 1/163)$ است. در نتیجه می‌توان نوشت

$$WGMSE^{nbinom} = (0/995 * 0/105) + (0/005 * 0/068) = 0/1048$$

چنان‌که دیده می‌شود براساس هر دو توزیع دوجمله‌ای منفی که جای‌گزین دو توزیع پواسون شده است مقدار مقایسه‌ای آماره‌های آزمون کوچک‌تر شده است. در واقع می‌توان نوشت:

کارایی ای‌آمیزه دوجمله ای منفی نسبت به ای‌آمیزه توزیع پواسون $= \frac{0/564}{0/1048} = 5/42$

بررسی انجام شده برای داده‌های مربوط به آزمون دکتری بیان می‌کند که استفاده از آمیزه‌ای متناهی از توزیع‌های دوجمله‌ای منفی به مراتب برتری چشم‌گیری نسبت به استفاده از آمیزه‌ای متناهی از دو توزیع پواسون دارد. متغیر توانایی نیز در هر دو جامعه (سال ۹۲ و ۹۳) به‌عنوان متغیراثرگذار تعیین می‌شود. این موضوع در پژوهش شبیه‌سازی نیز تأیید شد.

تشکر و قدردانی

از نکات ارزنده‌ای که داوران محترم ارایه کردند، کمال تشکر را داریم.

منابع

1. Chen and Holland, "New Equating Methods and Their Relationships With Levine Observed Score Linear Equating Under The Kernel Equating Framework." Psychometrika, Vol. 75, NO. 3 (2010) 542–557.
2. Cho H., Fryzlewicz P.,"High-Dimensional Variable Selection via Tilting", J. Roy. Stat. Soc. Ser. B, Stat. Method., 74 (2012) 593-622.
3. Du Y., Khalili A., Neslehova J. G., Steele R. J., "Simultaneous Fixed and Random Effects Selection in Finite Mixture of Linear Mixed-Effects Models", The Canadian Journal of Statistics, 41 (2013) 596-616.
4. Eskandari F., Ormoz E., "Finite Mixture of Generalized Semiparametric Models: Variable Selection via Penalized Estimation", Communications in Statistics–Simulation and Computation,(to appear) (2016).
5. Fan J., Li R., "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties", J. Amer. Statist. Assoc., 96 (2001)1348-1360.

6. Keller L. A., Skorupski W. P., Swaminathan H., Jodoin M. G. An Evaluation of Item Response Theory Equating Procedures for Capturing Changes in Examinee Distributions with Mixed-Format Tests. Paper Presented at the Annual Meeting of the National Council on Measurement in Education (2004).
7. Khalili A., Chen J., "Variable Selection in Finite Mixture of Regression Models", Journal of American Statistical Association, 102 (2007)1025-1038.
8. Khalili A., "An Overview of the New Feature Selection Methods in Finite Mixture of Regression Mdels", JIRSS, 10 (2011) 201-235.
9. Kolen M. J., Brennan R. L., "Test equating, scaling, and linking: Methods and 31 Practices", New York: Springer (2004).
10. Li R., Liang H., "Variable Selection in Semiparametric Regression Modeling", Ann. Statist., 36 (2008) 261-286.
11. Lee W., Ban J. "A comparison of IRT Linking Procedures. Applied Measurement in Education, 23 (2010) 23-48.
12. Ma S., Song Q., Wang L., "Simultaneous Variable Selection and Estimation in Semiparametric Modeling of Longitudinal/Clustered Data", Bernoulli, 19 (2013) 252-274.
13. McLachlan G. J., Peel D., "Finite Mixture Models", New York: Wiley (2000).
14. Nelder J., Wedderburn R. W. M, "Generalized Linear Models", J. Roy. Statist. Soc. Ser. A., 135 (1972) 370-384.
15. Myint A., Htet L., O., "An Application of Linear Test Equating Method in Scoring", Yangon Institute of Education Research Vol. 2, No. 1 (2010) 1-16.
16. Ormoz E., Eskandari F., "Variable Selection in Finite Mixture of Semi-Parametric Regression models", Commun Stat-Theorm, to appear (2013).
17. Ormoz E., Eskandari F., "Variable Selection in Finite Mixture of Semi-Parametric Regression Models", Communications in Statistics-Theory and Methods,Vol. 3 (2016)657-670.
18. Santarelli M. F., Latta D. D., Michele Scipioni M., PositanoV., Landini L., "A Conway-Maxwell–Poisson (CMP) model to address data dispersion on positron emission tomography, Computers in Biology and Medicine **77** (2016) 90-101.
19. Skrondal A., Rabe-Hesketh S., "Generalized Latent Variable Modelling: Multilevel, Longitudinal, and Structural Equations Models", Chapman and Hall (2004).
20. Schwarz G., "Estimating the Dimension of a Model", The Annals of Statistics, 6 (1978) 461-464.

21. Von Davier A. A., Holland P. W., Thayer D. T. "The kernel Method of Test Equating",New York: Springer-Verlag (2004).

22. Von Davier A. A. "A Statistical Perspective on Equating Test Scores", Scaling and Linking (pp.1-17). New York, NY: Springer-Verlag (2011).

23. Santarelli M. F., Latta D. D., Michele Scipioni M., Positano V., Landini L., "A Conway–Maxwell–Poisson (CMP) model to address data dispersion on positron emission tomography, Computers in Biology and Medicine **77** (2016) 90-101.