https://doi.org/ 10.22034/jr_iss.2024.732795
اندیشه آماری، سال ۱۴۰۳، دوره ۲۹، شماره ۱، ص ۱۴۷-۱۳۳

مدل‌های تلسکوپی آماسیده در صفر و کاربردهای آن

هادی صبوری[۱]* و مهدی دوست‌پرست[۲]

[۱] گروه آمار، دانشگاه زابل، زابل، ایران
[۲] گروه آمار، دانشگاه فردوسی مشهد، ایران

تاریخ دریافت: ۱۴۰۴/۰۷/۰۵
تاریخ پذیرش: ۱۴۰۴/۰۹/۲۹

چکیده:

داده‌های گسسته آماسیده در عمل، به طور گسترده مورد استفاده قرار می‌گیرند. یکی از ضروری‌ترین روش‌های مدل‌سازی برای این داده‌ها استفاده از مدل‌های مبتنی بر توزیع‌های آماسیده است. از آنجا که انتخاب توزیع پایه نقش اساسی در تعریف یک خانواده از توزیع‌های آماسیده دارد، بنابراین کارایی و توانایی برازش آن مستقیماً با توزیع پایه مرتبط است. در بیشتر تحقیقات مربوط به داده‌های آماسیده، مدل‌هایی مبتنی بر توزیع پواسون استفاده می‌شود. توزیع پواسون توزیع بسیار قدرتمندی است، اما یک ویژگی دارد که همان ویژگی به نقطه ضعف این توزیع در کاربرد بدل می‌شود. این ویژگی برابری واریانس و میانگین است. معمولا این شرط در عمل برای داده‌های واقعی به ندرت برقرار است. پس باید به دنبال یک جایگزین برای توزیع پواسون بود. چه بهتر که این جایگزین یک خانواده گسترده از توزیع‌ها با قابلیت‌های بیشتر باشد. یکی از این گزینه‌ها، خانواده توزیع‌های تلسکوپی است.

در این تحقیق، از خانواده توزیع‌های تلسکوپی به عنوان توزیع پایه برای تعریف یک رده عمومی از مدل‌های گسسته آماسیده استفاده می‌شود. این خانواده شامل توزیع‌هایی است که کمتر به عنوان توزیع پایه استفاده شده‌اند (مانند توزیع وایبول گسسته). از طرفی این خانواده به عنوان توزیع طول عمر نیز استفاده می‌شوند. همچنین به دلیل ارتباط با خانواده‌ای مهم از توزیع‌های پیوسته، طیف گسترده‌ای از سایر توزیع‌های آماری را نیز شامل می‌شود. وسعت این خانواده یک مزیت بزرگ برای آن محسوب می‌شود. بنابراین، اگر از خانواده تلسکوپی به عنوان توزیع پایه برای تعریف مدل‌های آماسیده استفاده شود، مدل‌های گسترده‌ای برای برازش هر مجموعه داده گسسته آماسیده (و حتی غیر آماسیده) وجود خواهد داشت. علاوه بر این در این تحقیق به مدل رگرسیونی مبتنی بر پارامتر آمیختگی پرداخته شده است.

واژه‌های کلیدی: داده‌های آماسیده، توزیع‌های آماسیده، مدل‌های خطی تعمیم‌یافته، خانواده توزیع‌های تلسکوپی.

۱ مقدمه

مدل‌های آماسیده بر اساس یک توزیع احتمالی آماسیده تعریف می‌شوند. این توزیع آماسیده بر اساس یک توزیع گسسته متداول آماری است (البته این مطلب قابل ذکر است که اگر داده‌های آماسیده پیوسته باشند در اینصورت توزیع پایه نیز پیوسته در نظر گرفته می‌شود). این مدل‌ها به طور گسترده در ادبیات علمی مورد استفاده قرار گرفته‌اند. این مدل‌ها اولین بار توسط کاتی و همکاران [۱۷] پایه‌گذاری شدند. آنها معیاری برای بررسی انعطاف‌پذیری توزیع‌ها استفاده کردند و نشان دادند که توزیعی به نام توزیع پواسون لگ–صفر[۱] وجود دارد که از توزیع‌های پواسون، دو جمله‌ای منفی و دو جمله‌ای انعطاف‌پذیرتر است. مهم‌ترین نقطه شروع برای مدل‌های آماسیده، توزیع پواسون آماسیده در صفر (*ZIP*) است که توسط لمبرت [۲۲] معرفی شد. برای مطالعه اثرات متغیرهای همراه. لمبرت یک مدل رگرسیون ZIP پارامتری با پیش‌بین‌های خطی از طریق تابع پیوند مناسب پیشنهاد کرد و آن را برای تحلیل ایرادهای لحیم‌کاری روی بردهای مدار چاپی به کار برد. مدل‌های رگرسیون *ZIP* بعدها با موفقیت در کاربردهای مهم مختلفی استفاده شده است. واندنبروک [۴۱] یک مدل رگرسیون پواسون استاندارد را به عنوان جایگزین مدل *ZIP* معرفی کرد. جانساکول و هایند [۱۵] این مدل را گسترش دادند که در آن احتمال صفر ممکن است با برخی

*نویسنده مسئول: saboori_hadi@yahoo.com

[1] log–zero Poisson distribution

متغیرهای همراه وابستگی داشته باشد. ریداوت و همکاران [۳۵] یک آزمون برای مقایسه مدل رگرسیون *ZIP* در برابر مدل دو جمله‌ای منفی آماسیده در صفر (*ZINB*) معرفی کردند. جانگ و همکاران [۱۶] یک روش بوت‌استرپ پارامتری پیشنهاد کردند. آگاروال و همکاران [۱] مدل رگرسیون *ZIP* را روی داده‌های شمارشی فضایی اعمال کردند. خوش‌گفتار و همکاران [۱۸] مدل‌های رگرسیون *ZIP* را برای برآورد کیفیت و کارایی یک نرم‌افزار به کار بردند. آن‌ها مشاهده کردند که مدل رگرسیون *ZIP* برای پیش‌بینی قوی‌تر از مدل رگرسیون پواسون عمل می‌کند. مدل رگرسیون *ZIP* اغلب در تحقیقات بهداشت عمومی برای بررسی رابطه بین متغیرهای مورد علاقه و یک خروجی گسسته که صفرهای زیادی دارد، استفاده می‌شود. لانگ و همکاران [۲۴] یک مدل رگرسیونی تحت عنوان *ZIP* حاشیه‌ای را توسعه دادند. ژو و همکاران [۴۵] مدل‌های شمارش آماسیده در صفر را برای اثرات تصادفی ناهمگون با مدل‌سازی واریانس آن‌ها به عنوان تابعی از متغیرهای همراه گسترش دادند. لیم و همکاران [۲۳] یک مدل رگرسیون *ZIP* آمیخته را برای داده‌هایی که هم صفرهای اضافی و هم پراکندگی بیش از حد ناشی از ناهمگونی مشاهده نشده داشتند را پیشنهاد کردند. تحلیل‌های رگرسیون با تکنیک‌های بیزی توسط گوش و همکاران [۱۳]، چن [۷]، داگنه [۱۰] و موسیو و همکاران [۲۹] مورد بررسی قرارگرفت. عامل پنهان بیزی در مدل *ZIP* توسط نیلون و چونگ [۳۱] برای تحلیل تفاوت‌های مولکولی بین بیماران سرطان پستان پیشنهاد شد.

در بیشتر مقالات مرتبط، توزیع‌های استفاده شده در مدل‌های آماسیده، متعلق به خانواده توزیع‌های سری توانی هستند. اما خانواده‌های مهم دیگری در توزیع‌های گسسته وجود دارند. یکی از این خانواده‌های مهم، خانواده توزیع‌های تلسکوپی است که توسط رضایی رکن‌آبادی و همکاران [۳۶] پیشنهاد شده است. این خانواده شامل توزیع‌های مهمی است که کمتر به عنوان توزیع‌های پایه در مدل‌های آماسیده استفاده شده‌اند. علاوه بر این، ارتباط این مدل با توزیع‌های پیوسته (قضیه ۱)، به محقق اجازه می‌دهد تا بهترین مدل را برای داده‌ها در طیف گسترده‌ای از توزیع‌ها انتخاب کند. بنابراین، ویژگی مهم این خانواده در مقایسه با مدل‌های آماسیده معمول (سری توانی)، وسعت بیشتر و معرفی مدل‌های جدیدتر با قابلیت‌های متفاوت‌تر است.

در این مقاله، مدل‌های مبتنی بر توزیع‌های آماسیده در صفر با استفاده از خانواده توزیع‌های تلسکوپی (*ZIT*) معرفی می‌شود. برای این هدف، در بخش ۲، توزیع‌های تلسکوپی و توزیع‌های آماسیده متناظر با آن را معرفی می‌کنیم. همچنین رابطه این خانواده از توزیع‌ها با توزیع‌های پیوسته ارائه می‌شود. در بخش ۳، مطالبی را درباره نحوه برآورد پارامترهای مدل معرفی شده ارائه داده و مدل رگرسیون این خانواده از توزیع‌ها را در بخش ۴ معرفی می‌کنیم. در بخش‌های ۵ و ۶، به آزمون فرضیه روی پارامترهای مدل رگرسیون و معرفی باقیمانده‌های مدل می‌پردازند. در بخش‌های ۷ و ۸، کارایی مدل معرفی شده را با مطالعه شبیه‌سازی و یک مجموعه داده واقعی ارزیابی می‌کنیم. در پایان در بخش ۹، یک خلاصه کلی و نتیجه‌گیری را ارائه می‌دهیم.

۲ خانواده توزیع‌های ZIT

خانواده توزیع‌های تلسکوپی توسط رضایی رکن‌آبادی و همکاران (۲۰۰۹) معرفی شد. این خانواده از توزیع‌ها به صورت زیر تعریف می‌شود:

تعریف ۱: یک متغیر تصادفی گسسته غیرمنفی Y دارای توزیع تلسکوپی است، با $Y \sim T(q, m_\theta)$ نشان داده می‌شود و تابع جرم احتمال آن به شکل زیر است:

$$g(y;q,m_\theta) = q^{m_\theta(y)} - q^{m_\theta(y+1)}, \quad y = 0,1,2,\cdots, \qquad (1)$$

که در آن $0 < q < 1$، θ یک بردار از پارامترها و $m_\theta(y)$ یک تابع اکیدا صعودی از y به طوریکه $m_\theta(0) = 0$ و $m_\theta(x) \to \infty$ وقتی $x \to \infty$، است.

به راحتی می‌توان نشان داد که $\sum_{y=0}^{\infty} g(y;q,m_\theta) = 1$، در نتیجه $g(y;q,m_\theta)$ در معادله (۱) در واقع یک تابع جرم احتمال است. خانواده‌های توزیع هندسی، رایلی گسسته و وایبول گسسته متعلق به خانواده تلسکوپی هستند. علاوه بر این هر عضو این خانواده یک توزیع پیوسته مرتبط با ویژگی‌های مشابه دارد. به عنوان یک حالت خاص فرض کنید $m_\theta(y) = \theta y$، $\theta > 0$، در اینصورت از (۱) توزیع هندسی نتیجه می‌شود، زیرا:

(۲)

$$g(y;q,m_\theta) = q^{m_\theta(y)} - q^{m_\theta(y+1)} = q^{\theta y}(1-q^\theta), \quad y = 0,1,2,\cdots.$$

این خانواده از توزیع‌ها، علاوه بر یک سری توزیع‌های گسسته مهم (جدول ۱)، به یک خانواده مهم از توزیع‌های پیوسته نیز مرتبط است. در ادامه، این ویژگی را بررسی کرده و اهمیت این ارتباط را در بخش‌های عملی تحلیل خواهیم کرد. یکی از مهم‌ترین نتایج این رابطه آن است که

جدول ۱: خانواده توزیع‌های تلسکوپی

توزیع	$m_\theta(y)$	پارامتر	مرجع
Geometri (GEO)	y	$\theta=1$	-
Discrete Rayleigh (DRA)	y^2	$\theta=1$	[۳۷]
Discrete Weibull (DW)	y^α	$\theta=\alpha$	[۳۰]
Discrete modified Weibull (DMW)	$y^\alpha\gamma^y$	$\theta=(\alpha,\gamma)$	[۳۳]
Discrete reduced modified Weibull (DRMW)	$\sqrt{y}(1+\alpha\gamma^y)$	$\theta=(\alpha,\gamma)$	[۴]

جدول ۲: توزیع‌های پیوسته مهم مرتبط با خانواده توزیع‌های تلسکوپی بر مبنای قضیه (۱)

توزیع	$-\alpha m_\theta(y)$	پارامتر	مرجع
Exponential (E)	$-\lambda y$	$\alpha=\lambda,\ \theta=1$	-
Weibull (W)	$-\left[\frac{y}{\alpha}\right]^\beta$	$\alpha=1,\ \theta=\alpha$	-
Rayleigh (RA)	$-\frac{y^2}{2\alpha^2}$	$\alpha=1,\ \theta=(\alpha,\beta)$	-
Linear Exponential (LE)	$-\left[\alpha y+\frac{\beta}{2}y^2\right]$	$\alpha=1,\ \theta=(\alpha,\beta)$	-
Modified Weibull (MW)	$-\beta y^\gamma\exp(\lambda y)$	$\alpha=\beta,\ \theta=(\alpha,\gamma)$	[۲۱]
Almalki and Yuan's modified Weibull (AYMW)	$-\left[\alpha y^\theta+\beta y^\gamma\exp(\lambda y)\right]$	$\alpha=1,\ \theta=(\alpha,\beta,\gamma,\lambda,\theta)$	[۳]
New Modified Weibull (NMW)	$-\left[\alpha y^\theta+\beta y^\gamma\right]$	$\alpha=1,\ \theta=(\alpha,\beta,\gamma,\theta)$	[۴۴]
Sarhan and Zaindin's modified Weibull (SZMW)	$-\left[\alpha y+\beta y^\gamma\right]$	$\alpha=1,\ \theta=(\alpha,\beta,\gamma)$	[۴۰]
Chen	$-\lambda[\exp(y^\beta)-1]$	$\alpha=\lambda,\ \theta=\beta$	[۸]
Exponential Extension (EE)	$-[(1+\lambda y)^\alpha-1]$	$\alpha=1,\ \theta=(\alpha,\lambda)$	[۳۲]
Exponential Powe (EP)	$-\{\exp[(\lambda y)^\alpha]-1\}$	$\alpha=1,\ \theta=(\alpha,\lambda)$	[۷]
Gompertz (GOM)	$\frac{-\theta}{\alpha}[exp(-\alpha y)-1]$	$\alpha=\theta,\ \theta=\alpha$	[۲۷]
Generalized power Weibull (GPW)	$-[(1+y^\alpha)^\theta-1]$	$\alpha=1,\ \theta=(\alpha,\theta)$	[۳۲]
Pham	$-[\lambda y^\alpha-1]$	$\alpha=1,\ \theta=(\alpha,\lambda)$	[۳۴]
Linear failure rate (LFR)	$-\left[\alpha y+\frac{\beta y^2}{2}\right]$	$\alpha=1,\ \theta=(\alpha,\beta)$	[۵]
log-gamma (LG)	$-\alpha\exp(\lambda y)$	$\alpha=\alpha,\ \theta=\lambda$	[۱۹]
Weibull Extension (WEE)	$-\alpha\left[\exp\left(\frac{y}{\alpha}\right)^\beta-1\right]$	$\alpha=\alpha,\ \theta=(\alpha,\beta)$	[۲۸]
BurrX (BX)	$-[\lambda y]^2$	$\alpha=1,\ \theta=\lambda$	[۲۰]

طیف گسترده‌ای از توزیع‌ها برای مدل‌سازی داده‌ها، با برخی مشاهدات آماسیده، در دسترس است.

فرض کنید W یک متغیر تصادفی پیوسته غیرمنفی با تابع توزیع زیر باشد:

$$H(w;\alpha,m_\theta)=1-\exp[-\alpha m_\theta(w)],\quad w>0, \qquad (3)$$

که در آن $\alpha>0$ و θ یک بردار پارامتر (که ممکن است شامل α نیز باشد) و m_θ مشابه شکل تابع جرم احتمالی (۱) است.

تابع چگالی W به صورت زیر است:

$$h(w;\alpha,m_\theta)=\alpha m'_\theta(w)\exp[-\alpha m_\theta(w)],\quad w>0, \qquad (4)$$

که در آن

$$m'_\theta(w)=\frac{\partial m_\theta(w)}{\partial w}.$$

توزیع‌هایی با شکل تابع چگالی احتمالی (۴)، متعلق به خانواده توزیع‌های نمایی تعمیم‌یافته[۲] (EE) هستند([۱۴]). این خانواده شامل توزیع‌های مهمی مانند نمایی، رایلی، وایبول، خطی–نمایی، گومپرتز، وایبول اصلاح‌شده است.

خانواده توزیع‌های گسسته تلسکوپی با خانواده توزیع‌های EE مرتبط است، و این ارتباط با استفاده از قضیه (۱) بیان می‌شود.

قضیه ۱. فرض کنید W یک متغیر تصادفی پیوسته با توزیعی به شکل (۳) باشد و $Y=[W]$ (که در آن $[a]$ به معنای جزء صحیح a است)، در اینصورت $Y\sim T(q,m_\theta)$.

اثبات: برای اثبات قضیه به صورت زیر عمل می‌شود

$$\begin{aligned} g(y) &= P_Y(y)=\int_y^{y+1} dH(w;\alpha,m_\theta)=H(y+1)-H(y)\\ &= q^{m_\theta(y)}-q^{m_\theta(y+1)}=g(y;q,m_\theta),\quad y=0,1,2,\cdots, \end{aligned}$$

که در آن $q=\exp(-\alpha)$ است. برای جزئیات بیشتر، به رضایی رکن‌آبادی و همکاران [۳۶] مراجعه کنید. توزیع‌های پیوسته مرتبط با خانواده تلسکوپی با استفاده از قضیه ۱ در جدول ۲ ارائه شده‌اند.

تعریف ۲. خانواده توزیع‌های آماسیده در صفر با توزیع پایه (۱) دارای تابع جرم احتمال زیر است

$$f(Y;w,q,m_\theta)=\begin{cases} w+\eta(w)g(0;q,m_\theta), & y_i=0\\ \eta(w)g(y;q,m_\theta), & y>0, \end{cases} \qquad (5)$$

که در آن $\eta(w)=1-w$، $0\le\eta(w)\le 1$، و $g(\cdot;q,m_\theta)$ متعلق به خانواده تلسکوپی است. می‌گوییم Y دارای توزیع ZIT است و به صورت $Y\sim \mathrm{ZIT}(w,q,m_\theta)$ نشان می‌دهیم.

به عنوان مثال اگر توزیع هندسی را به عنوان توزیع پایه در مدل‌های آماسیده انتخاب کنیم، آنگاه توزیع هندسی آماسیده به دست می‌آید (ژیاوو و همکاران [۴۳]). اگر $Y\sim \mathrm{ZIT}(w,q,m_\theta)$ آنگاه

$$\begin{aligned} E(Y) &= \eta(w)E(X),\\ E(Y^2) &= \eta(w)E(X^2),\\ M_Y(t) &= E[\exp(ty)]=\eta(w)M_X(t), \end{aligned} \qquad (6)$$

که در آن متغیر تصادفی X متعلق به توزیع تلسکوپی است.

۳ برآورد

فرض کنید $\mathbf{Y}=(Y_1,\cdots,Y_n)$ یک نمونه تصادفی از $\mathrm{ZIT}(w,q,m_\theta)$ با تابع چگالی احتمال (۵) باشد، و $\mathbf{y}=(y_1,\cdots,y_n)$ بردار مشاهدات باشد. تابع درست‌نمایی (LF)، $\mathbf{y}$ به صورت زیر است

$$\begin{aligned} &L(\mathbf{y};w,q,m_\theta) \qquad (7)\\ &=\prod_{j=1}^{n}\Big\{[w+\eta(w)g(y_j;q,m_\theta)]^{I(y_j)}\\ &\times[\eta(w)g(y_j;q,m_\theta)]^{I(y_j>0)}\Big\}, \end{aligned}$$

که در آن

$$I(y)=\begin{cases} 1, & y=0\\ 0, & y\neq 0, \end{cases} \qquad (8)$$

و $I(y>0)=1-I(y)$. لگاریتم LF (۷)، (LLF) به صورت زیر است

$$\begin{aligned} \ell(w,q,m_\theta;\mathbf{y}) &= \log[L(\mathbf{y};w,q,m_\theta] \qquad (9)\\ &= \sum_{j=1}^{n}\Big\{I(y_j)\log\left(w+\eta(w)g(y_j;q,m_\theta)\right)\\ &+ I(y_j>0)\left[\log\eta(w)+\log g(y_j;q,m_\theta)\right]\Big\}. \end{aligned}$$

برآوردگرهای درست‌نمایی ماکزیمم (MLEs) پارامترها، با بیشینه‌سازی (۷) یا (۹) به دست می‌آیند. معادلات درست‌نمایی از (۹) به صورت زیر قابل حصول است

[2] Extended Exponential Distribution

$$\frac{\partial \ell(w,q,m_\theta;\mathbf{y})}{\partial w} = \sum_{j=1}^{n}\left\{I(y_j)\times\frac{1-g(y_j;q,m_\theta)}{w+\eta(w)g(y_j;q,m_\theta)}-\frac{I(y_j>0)}{\eta(w)}\right\}=0, \quad (10)$$

$$\frac{\partial \ell(w,q,m_\theta;\mathbf{y})}{\partial \theta} = \sum_{j=1}^{n}\left\{I(y_j)\times\frac{\eta(w)\partial g(y_j;q,m_\theta)/\partial\theta}{w+\eta(w)g(y_j;q,m_\theta)}+I(y_j>0)\times\frac{\partial g(y_j;q,m_\theta)/\partial\theta}{g(y_j;q,m_\theta)}\right\}=0, \quad (11)$$

$$\frac{\partial \ell(w,q,m_\theta;\mathbf{y})}{\partial q} = \sum_{j=1}^{n}\left\{I(y_j)\times\frac{\eta(w)\partial g(y_j;q,m_\theta)/\partial q}{w+\eta(w)g(y_j;q,m_\theta)}+I(y_j>0)\times\frac{\partial g(y_j;q,m_\theta)/\partial q}{g(y_j;q,m_\theta)}\right\}=0. \quad (12)$$

از آنجا که صورت بسته‌ای برای معادلات (۱۰) ، (۱۱) و (۱۲) وجود ندارد، بنابراین باید این معادلات را با روش‌های عددی تکراری حل کنیم. همانطور که می‌دانیم یکی از روش‌های متداول، برای بررسی دقت برآوردهای به دست آمده از روش‌های عددی تکراری، و فاصله اطمینان برای آنها، و انجام آزمون‌های آماری روی آنها، استفاده از خطای استاندارد آن برآوردها است. برای این منظور، باید ماتریس اطلاعات فیشر محاسبه شود. برای به دست آوردن ماتریس اطلاعات فیشر، باید مشتقهای مرتبه دوم لگاریتم تابع درست‌نمایی را به دست آوریم (پیوست).

۴ مدل رگرسیون ZIT

فرض کنید متغیرهای همراه $\mathbf{x}^\top=(1,x_1,\ldots,x_p)^\top$، $p\geq 1$، مرتبط با متغیر پاسخ Y وجود دارند. همچنین فرض کنید

$$Y\sim ZIT(w,q,m_\theta),$$

بنابراین، مدل رگرسیون با ۳ پارامتر (w,q,θ) داریم (البته اگر θ بردار باشد، به تناسب آن تعداد پارامترهای مدل افزایش پیدا می‌کند). این توزیع، یک توزیع آمیخته از ۲ توزیع است. اولین توزیع، یک توزیع تباهیده در صفر با وزن w و توزیع دوم توزیع تلسکوپی با وزن $\eta(w)$ است. فرض کنید $\mathbf{y}=(y_1,\cdots,y_n)$ یک نمونه تصادفی از توزیع $\mathrm{ZIT}(w,q,m_\theta)$ باشد. هر y_i با یک بردار از متغیرهای همراه مرتبط است که با $\mathbf{x}^\top=(1,x_1,\ldots,x_p)^\top$، $p\geq 1$ نشان داده شده است.

برای پیاده‌سازی مدل رگرسیون، با رویکردی مشابه اگرستی [۲] در مدل‌های خطی تعمیم‌یافته عمل می‌کنیم. ابتدا، یک یا چند پارامتر از مدل را برای رابطه بین متغیرهای همراه و متغیر پاسخ انتخاب می‌کنیم. اگر پارامترهای مورد نظر متعلق به توزیع پایه (q,θ) باشند، از یک تابع پیوند متناسب استفاده می‌کنیم، در مواردی که پارامتر مورد مطالعه، وزن w باشد، و با توجه به اینکه توزیع مورد نظر یک توزیع آمیخته است، از روش‌های عمومی مرتبط با مدل‌های خطی تعمیم‌یافته که برای مدل‌های مبتنی بر توزیع چندجمله‌ای توسعه یافته‌اند، استفاده می‌کنیم. این توزیع آمیخته متشکل از دو جزء است که معادل دو رده در نظر گرفته می‌شوند: یک توزیع تباهیده در نقطه صفر و یک توزیع تلسکوپی. این اجزا به ترتیب با وزن‌های w و $\eta(w)$ در ترکیب نهایی مشارکت می‌کنند (به عبارت دیگر از مدل متداول لجیت استفاده می‌کنیم). اکنون فرض کنید:

$$\gamma=\log\left(\frac{w}{\eta(w)}\right), \quad (13)$$

که در آن $\gamma=\mathbf{x}^\top\beta$، و $\mathbf{x}=(1,x_1,\ldots,x_p)^\top$ و p تعداد متغیرهای همراه است.

در این مدل، از توزیع تلسکوپی برای پاسخ Y استفاده می‌کنیم. به عنوان مثال، فرض کنید پارامترهای توزیع پایه معلوم باشند، بنابراین مدل رگرسیون $p+1$ پارامتر مجهول دارد. تعریف مدل رگرسیون را می‌توان به صورت زیر خلاصه کرد

$$\begin{cases} Y_j & \overset{id}{\sim} \ \mathrm{ZKIT}(w,q_j,m_{\theta_j}), \quad j=1,\ldots,n, \\ w & = \frac{\exp(\mathbf{x}^\top\beta)}{1+\exp(\mathbf{x}^\top\beta)}, \\ \eta(w) & = \frac{1}{1+\exp(\mathbf{x}^\top\beta)}, \\ \theta \ \text{ or } \ q & = h(\mathbf{v},\xi), \end{cases} \quad (14)$$

که در آن، $\beta=(\beta_0,\ldots,\beta_p)^\top$ و $\xi=(\xi_0,\ldots,\xi_s)^\top$ بردار پارامترهای مدل‌های رگرسیون هستند و $\mathbf{v}=(1,v_1,\ldots,v_s)^\top$ بردار متغیرهای همراه برای مدل رگرسیون مبتنی بر پارامترهای توزیع پایه است (s تعداد متغیرهای همراه است). همچنین در این مدل، $\eta(w)$ در (۱۴) به عنوان جمله پایه در مدل لوجیت چندجمله‌ای در نظر گرفته می‌شود. بنابراین، تابع لجیت مرتبط با $\eta(w)$ یک تابع خطی از مؤلفه‌های بردار متغیرهای

همراه $\mathbf{x}_j$ است به طوریکه:

$$\log\left(\frac{w}{\eta(w)}\right)=\mathbf{x}^\top\beta.$$

در (۱۴)، تابع پیوند $h(\mathbf{v},\xi)$ بسته به شکل توزیع پایه تعریف می‌شود. به عنوان مثال در توزیع وایبول گسسته اگر θ پارامتر توزیع پایه در نظر گرفته شود آنگاه تابع پیوند را به صورت $h(\mathbf{v},\xi)=\exp(\mathbf{v}^\top\xi)$ می‌توان تعریف کرد.

۱.۴ برآورد پارامترهای مدل رگرسیون ZIT

در این بخش، دو روش برای برآورد پارامترهای مجهول مدل ارائه می‌شود؛ یکی روش درست‌نمایی ماکزیمم و دیگری الگوریتم امید–ماکزیمم‌سازی[3] (EM). برای راحتی کار فرض می‌کنیم پارامترهای توزیع پایه معلوم هستند. اگر این پارامترها نیز مجهول فرض شوند، بسته به شکل توزیع پایه، LF متناظر را می‌توان بازنویسی و از روش‌های مشابه برای برآورد پارامترهای متناظر استفاده کرد. با فرض توزیع پایه تلسکوپی و جایگزینی (۱۳) در (۷)، LLF مدل رگرسیون ZIT را می‌توان از (۱۳) به صورت زیر محاسبه کرد:

$$\begin{aligned}&\ell_{obs}(\beta;\mathbf{y})\\&=\log L_{obs}(\beta;\mathbf{y})\\&\propto\Big\{I(y_j)\log\left[\exp(\gamma)+g(y_j;q,m_\theta)\right]-\log(1+\exp(\gamma))\Big\},\end{aligned}\tag{۱۵}$$

که در آن برآورد پارامترهای $\gamma=\mathbf{x}^\top\beta$ با ماکزیمم‌سازی (۱۵) نسبت به پارامترهای مجهول قابل محاسبه است. معادلات درست‌نمایی از (۱۵) به صورت زیر به دست می‌آیند:

$$\begin{aligned}&\partial\ell_{obs}(\beta;\mathbf{y})/\partial\beta\\&=\mathbf{x}^\top\exp(\gamma)\sum_{j=1}^{n}\left(\frac{I(y_j)}{\exp(\gamma)+g(y_j;q,m_\theta)}-\frac{1}{1+\exp(\gamma)}\right)=0.\end{aligned}$$

وقتی معادلات درست‌نمایی بسیار پیچیده است یکی از روش‌هایی که معمولاً جایگزین MLE می‌شود، استفاده از الگوریتم EM است. در رویکرد EM، داده‌های مشاهده شده $\mathbf{y}=(y_1,\cdots,y_n)$ را به عنوان بخشی از داده‌های کامل در نظر می‌گیریم، که شامل داده‌های گمشده، $\mathbf{z}=(\mathbf{z}_1,\cdots,\mathbf{z}_n)$ نیز می‌شود. در این حالت، $\mathbf{z}_j=(z_{j1},z_{j2}), j=1,\cdots,n$ یک بردار با ۲ مؤلفه و با تابع جرم احتمال زیر است

$$Pr\Big(\mathbf{z}_j=(z_{j1},z_{j2})\Big)=\begin{cases}w, & \mathbf{z}_j=(1,0)\\ \eta(w), & \mathbf{z}_j=(0,1).\end{cases}\tag{۱۶}$$

در واقع، توزیع متغیرهای پنهان $\mathbf{z}=(\mathbf{z}_1,\cdots,\mathbf{z}_n)$ یک توزیع دوجمله‌ای با پارامترهای $(1,w)$ است. تابع چگالی احتمال توام $(y_j,\mathbf{z}_j)$ به صورت زیر است

$$Pr(y_j,\mathbf{z}_j)=\begin{cases}w, & z_{j1}=1,y_j=0\\ \eta(w)\times g(y_j;q,m_\theta), & z_{j2}=1,y_j\geq 0.\end{cases}\tag{۱۷}$$

بنابراین

$$Pr\Big(Y_j=y_j|\mathbf{z}_j=(z_{j1},z_{j2})\Big)=\begin{cases}1, & z_{j1}=1,y_j=0\\ g(y_j;q,m_\theta), & z_{j2}=1,y_j\geq 0.\end{cases}\tag{۱۸}$$

و در نتیجه توزیع توام داده‌های کامل (داده‌های مشاهده شده به اضافه داده‌های گمشده) با استفاده از (۱۶) و (۱۸) به صورت زیر به دست می‌آید

$$\begin{aligned}&Pr\Big(Y_j=y_j,\mathbf{z}_j=(z_{j1},z_{j2})\Big)\\&=Pr\Big(Y_j=y_j|\mathbf{z}_j=(z_{j1},z_{j2})\Big)\times Pr\Big(\mathbf{z}_j=(z_{j1},z_{j2})\Big)\\&=\begin{cases}w, & z_1=1,y=0\\ \eta(w)\times g(y_j;q,m_\theta), & z_2=1,y\geq 0.\end{cases}\end{aligned}\tag{۱۹}$$

بنابراین، با استفاده از شکل تابع جرم احتمال به دست آمده در (۱۹) و رابطه (۱۵)، LF داده‌های کامل تحت مدل ZIT به صورت زیر حاصل می‌شود

$$\begin{aligned}&L_{comp}(w,q,\theta;\mathbf{y},\mathbf{z})\\&=\prod_{j=1}^{n}\Big\{\left[w^{z_{1j}\times I(y_j)}\right]\times\left(\eta(w)g(y_j;q,m_\theta)\right)^{z_{2}}\Big\},\end{aligned}\tag{۲۰}$$

با استفاده از (۱۳)، لگاریتم تابع درست‌نمایی داده‌های کامل $(\mathbf{y},\mathbf{z})$ را می‌توان به صورت زیر بازنویسی کرد

$$\begin{aligned}\ell_{comp}(w,q,\theta;\mathbf{y},\mathbf{z}) &= \log L_{comp}(w,q,,\theta;\mathbf{y},\mathbf{z})\\ &= \sum_{j=1}^{n}\Big\{z_{1j}I(y_j)\Big[\gamma-\log(1+e^{\gamma})\Big]\\ &+ z_{2j}\log\eta(w)+z_{2j}\log g(y_j;q,m_\theta)\Big\}.\end{aligned}\tag{۲۱}$$

اکنون الگوریتمی مشابه الگوریتم EM ارائه شده توسط وو (۱۹۸۳) برای مدل ZIT را توصیف می‌کنیم. اولین مرحله در الگوریتم EM انتخاب برخی مقادیر اولیه برای پارامترهای مجهول است. انتخاب مقادیر اولیه نقش مهمی در همگرایی الگوریتم EM دارد. انتخاب نامناسب مقادیر اولیه می‌تواند منجر به عدم همگرایی الگوریتم و در نتیجه منجر به عدم موفقیت در برآورد پارامترها شود. پیشنهاد ما استفاده از نسبت صفرها در مجموعه داده برای مقادیر اولیه در برآورد بردار پارامتر w است. از

[3] Expectation–maximization algorithm

جدول ۳: مقادیر $E(\mathbf{z}|\mathbf{y})$ برای مدل رگرسیونی ZIT

$\mathbf{z}$	$y=0$	$y\geq 0$
z_1	$\frac{\exp(\gamma)}{\exp(\gamma)+g(0;q,m_\theta)}$	0
z_2	$\frac{g(0;q,m_\theta)}{\exp(\gamma)+g(0;q,m_\theta)}$	1

معادله (۱۳)، مقدار اولیه γ_0 را برای γ انتخاب می‌کنیم. مرحله بعدی مقداردهی اولیه متغیرهای پنهان $\mathbf{z}$ با استفاده از مقادیر مورد انتظار آنها است که گام E نامیده می‌شود. از مقادیر امید ریاضی شرطی $E(\mathbf{z}|\mathbf{y})$ که در جدول ۳ داده شده‌اند استفاده می‌کنیم. بنابراین، گام E الگوریتم EM به صورت زیر است

$$\hat{z}_{1j} = E(z_{1j}|y_j=0) = \frac{\exp(\gamma)}{\exp(\gamma)+g(0;q_j,m_{\theta_j})}$$
$$\hat{z}_{1j} = E(z_{1j}|y_j\neq 0) = 0.$$

از (۲۱) لگاریتم تابع درستنمایی کامل برای برآورد بردار پارامترهای β، به صورت زیر است:

$$\ell_{comp}(\beta;\mathbf{y},\mathbf{z}) \propto \sum_{j=1}^{n}\left\{z_{1j}I(y_j)\left[\mathbf{x}^\top\beta-\log\left(1+\exp[\mathbf{x}^\top\beta]\right)\right] - z_{2j}\log\left(1+\exp[\mathbf{x}_l^\top\beta]\right)\right\}, \quad (۲۲)$$

برای مرحله بیشینه‌سازی الگوریتم EM، باید به جای بیشینه‌سازی لگاریتم LF کامل (۲۲)، معادلات امتیاز زیر را حل کنیم:

$$\frac{\partial \ell_{comp}}{\partial\beta} = \sum_{j=1}^{n}\left\{\hat{z}_{1j}I(y_j)\mathbf{x}-\frac{\mathbf{x}\exp[\mathbf{x}^\top\beta]}{1+\exp[\mathbf{x}^\top\beta]}\right\}. \quad (۲۳)$$

برای محاسبه خطاهای استاندارد برآوردهای به دست آمده از الگوریتم EM، می‌توان از رویکرد طراحی شده توسط لویس [۲۶] نیز استفاده کرد.

۵ آزمون فرضیه‌ها

هدف این بخش معرفی روشی برای انجام آزمون‌های مختلف روی پارامترهای مدل است. فرض کنید می‌خواهیم آزمون کنیم که آیا تورم در نقطه صفر رخ داده است یا خیر. این به معنای آزمون فرضیه صفر $H_0: w=0$ در برابر فرضیه جایگزین $H_1: w>0$ است. همچنین آزمون ضرورت متغیر همراه jام در مدل را می‌توان با استفاده از فرضیه $H_0: \beta_j=0$ در برابر $H_1: \beta_j\neq 0$ انجام داد $(j=0,1,\ldots,p)$. این آزمون‌ها را می‌توان با استفاده از آماره آزمون والد[4]، $z=\hat{\beta}_i/SE(\hat{\beta}_i)$ انجام داد. ثابت شده است که تحت فرضیه H_0، این آماره دارای توزیع نرمال استاندارد است، که در آن $SE(\hat{\beta}_j)$ خطای استاندارد (SE) برآوردگر $\hat{\beta}_j$ است. یادآوری می‌شود که یک روش جایگزین استفاده از آزمون‌های نسبت درستنمایی خطی تعمیم‌یافته[5] (GLRTs) است. برای مثال، آماره آزمون GLRT برای $H_0: \beta_j=0$ در مقابل $H_1: \beta_j\neq 0$ به صورت زیر است

$$-2\log\frac{L_{obs}(\widetilde{\beta},\beta_j=0)}{L_{obs}(\hat{\beta})}. \quad (۲۴)$$

این آماره به طور مجانبی دارای توزیع کای-دو با یک درجه آزادی است. $\hat{\beta}$، MLE بردار پارامتر β و $\widetilde{\beta}$، MLE β، تحت فرض صفر $H_0: \beta_j=0, j=0,\cdots,p$ است.

۶ تحلیل باقیمانده‌ها

یکی از مهم‌ترین روش‌ها برای بررسی کارایی یک مدل، تحلیل باقیمانده‌های آن است. در واقع، تحلیل باقیمانده‌های یک مدل نوعی بررسی ضربان قلب آن مدل خواهد بود. اما تحلیل باقیمانده‌های مدل در مجموعه داده‌های گسسته مشکلات زیادی دارد. چون به دلیل ماهیت گسسته چنین داده‌هایی، قاعدتا توزیع آن‌ها نرمال نیست. یکی از مهم‌ترین پیشنهادات برای حل این مشکل، استفاده از باقیمانده‌های

[4] Wald test
[5] Generalized Likelihood Ratio Test
[6] Randomized Quantile Residuals

چندکی تصادفی‌شده[۶] (RQR) است. PQR توسط فنگ و همکاران [۱۲] معرفی شد. آنها این باقیمانده‌ها را برای غلبه بر مشکلات استفاده از باقیمانده‌های معمول در مدل‌های مبتنی بر داده‌های گسسته معرفی کردند. آنها نشان دادند که PQR تحت مدل واقعی از یک توزیع نرمال پیروی می‌کند. PQR ها با معکوس کردن تابع توزیع تجمعی برازش‌شده برای هر پاسخ و یافتن چندک نرمال استاندارد متناظر به دست می‌آیند. فرض کنید $G(\mathbf{y};w,\theta)$ تابع توزیع تجمعی متغیر تصادفی پیوسته Y باشد، سپس $G(Y_i;w,\theta)$ از توزیع یکنواخت استاندارد پیروی می‌کند. PQR به صورت زیر تعریف می‌شود

$$q_i = \Phi^{-1}[G(y_i;\hat{w}_i,\hat{\theta}_i)], \qquad (۲۵)$$

که در آن $\Phi^{-1}(.)$ تابع چندک توزیع نرمال استاندارد است. حال، اگر گسسته باشد (چیزی که در این مطالعه با آن روبرو هستیم)، در این حالت یک تصحیح جزئی ضروری است:

$$G^*(y;w,\theta) = G(y^-;w,\theta) + U \times g(y;w,\theta),$$

که در آن $g(\mathbf{y};w,\theta)$ یک تابع روی تکیه‌گاه Y است و $G^*(\mathbf{y};w,\theta)$ به عنوان اصلاح‌شده Y تعریف می‌شود. همچنین U یک متغیر تصادفی با توزیع یکنواخت استاندارد است، و $G(\mathbf{y}^-;w,\theta)$ حد پایین G است. وقتی G، یک تابع توزیع گسسته است، a_i و b_i به صورت زیر تعریف می‌شوند

$$\begin{aligned} a_i &= lim_{y\to y_i^-} G(y;\hat{w}_i,\hat{\theta}_i), \\ b_i &= G(y_i;\hat{w}_i,\hat{\theta}_i), \end{aligned}$$

و در نهایت، PQR ها به صورت زیر تعریف می‌شوند

$$q_i = \Phi^{-1}[G_i^*],$$

که در آن G_i^* دارای توزیع یکنواخت روی بازه $(a_i,b_i]$ بوده و q_i نیز دارای توزیع نرمال استاندارد است. با توجه به رویکرد تعریف‌شده برای به دست آوردن PQR، واضح است که تنها نیاز به تعریف است. در ادامه، در بخش تحلیل مدل با استفاده از داده‌های واقعی، یکی از ابزارهایی که برای بررسی کارایی مدل استفاده می‌کنیم، PQR ها هستند.

۸ کاربرد

در این بخش، سعی می‌کنیم توانایی مدل‌های رگرسیونی مبتنی بر خانواده توزیع‌های تلسکوپی آماسیده را با مجموعه داده واقعی ارزیابی کنیم. تقریباً هیچ محدودیتی در انتخاب مجموعه داده (گسسته) نداریم. زیرا این خانواده شامل مدل‌های زیادی است که دست محقق را برای برازش هر مجموعه داده گسسته یا حتی پیوسته باز می‌گذارد.

برای مجموعه داده واقعی، از بین ۱۷ توزیع از خانواده توزیع‌های تلسکوپی آماسیده (علاوه بر این، سه توزیع مهم که به این خانواده تعلق

۷ مطالعه شبیه‌سازی

در این بخش، از یک مطالعه شبیه‌سازی استفاده شده است. این مطالعه برای بررسی اطمینان از عملکرد مدل و همچنین اطمینان از برآورد درست پارامترهای مدل آماسیده معرفی شده، ارائه شده است. برای این منظور سه زیرتوزیع از خانواده تلسکوپی انتخاب شده است: توزیع‌های هندسی، رایلی گسسته و وایبول گسسته. ابتدا نمونه‌هایی تصادفی با حجم‌های ۵۰، ۱۰۰، ۵۰۰ و ۱۰۰۰ از توزیع‌های هندسی، رایلی گسسته و وایبول گسسته آماسیده در صفر تولید شده، سپس خود آن توزیع‌ها و توزیع پواسون آماسیده در صفر (به عنوان متداول‌ترین مدل آماسیده در صفر) بر این مجموعه داده‌ها برازش داده شده است. این مراحل را ۱۰،۰۰۰ بار تکرار کرده و اریبی برآوردگرهای درستنمایی ماکزیمم و میانگین توان‌های دوم خطا ،(MSE) برای پارامترها، استخراج شده است. مقادیر پارامتر w، یعنی وزن نقطه آماسیده برای همه مدل‌ها ۰٫۱ درنظر گرفته شده و برای تمام توزیع‌ها پارامتر $q = ۰٫۱$، و درتوزیع وایبول گسسته پارامتر $\theta = ۰٫۵$ در نظر گرفته شده است.

نتایج این شبیه‌سازی‌ها در جدول ۴ نشان داده شده است. عدد اول اریبی برآورد پارامترهای توزیع از روش MLE است، و عدد داخل پرانتز MSE آن برآوردها است. با توجه به نتایج جدول ۴ و با استفاده از سه معیار $-\log(L)$، معیار اطلاعات آکائیک (AIC) و معیار اطلاعات بیزی (BIC) می‌توان دریافت که بهترین برآورد در هر مرحله از شبیه‌سازی متعلق به توزیع واقعی است (که داده از آن تولید شده است). همچنین، میزان اریبی در برآوردها و MSE با افزایش حجم نمونه تقریباً کاهش می‌یابد.

$$\begin{aligned} AIC &= ۲ \times s - ۲ \times LLF, \\ BIC &= s\log(n) - ۲ \times LLF, \end{aligned}$$

که در آن، s تعداد پارامترهای برآورد شده در مدل است.

جدول ۴: شبیه‌سازی در خانواده ZIT

توزیع	w	α	q	λ	$-\log(L)$	AIC	BIC
$n=۵۰$							
GEO	۰٫۳۲۸	–	−۰٫۰۱۲۴	–	۶۳٫۴۹۷۱	۱۲۸٫۹۹۴۲	۱۳۰٫۹۰۶۲
	(۰٫۰۲۳۴)	–	(۰٫۰۰۶۳)	–	–	–	–
$Poisson$	۰٫۳۲۶۶	–	–	۳٫۱۱۲۲	۶۳٫۴۲۵۳	۱۲۸٫۸۵۰۶	۱۳۰٫۷۶۲۶
	(۰٫۱۲۰۴)	–	–	–	–	–	–
DR	−۰٫۰۰۲۳	–	−۰٫۰۰۲۴	–	۸۴٫۲۹۲۷	۱۷۰٫۵۸۵۴	۱۷۲٫۴۹۷۴
	(۰٫۰۰۳۷)	–	(۰٫۰۰۰۴)	–	–	–	–
$Poisson$	۰٫۰۱۳۰	–	–	۳٫۰۷۷۷	۸۴٫۳۶۷۷	۱۷۰٫۷۳۵۴	۱۷۲٫۶۴۷۴
	(۰٫۰۰۴۴)	–	–	–	–	–	–
DW	۰٫۲۳۴۸	۰٫۲۶۸۶	۱٫۴۵۷۶	–	۶۳٫۳۴۶۵	۱۳۰٫۶۹۳	۱۳۴٫۵۱۷
	(۰٫۰۹۹۸)	(۰٫۱۱۵۱)	(۶٫۴۶۵۷)	–	–	–	–
$Poisson$	۰٫۳۲۵۴	–	–	۳٫۱۱۰۵	۶۳٫۶۸۷۵	۱۲۹٫۳۷۵	۱۳۱٫۲۸۷
	(۰٫۱۲۰۱)	–	–	–	–	–	–
$n=۱۰۰$							
GEO	۰٫۰۱۵۰	–	−۰٫۰۰۸۱	–	۱۲۹٫۰۳۱۵	۲۶۰٫۰۶۳۰	۲۶۲٫۶۶۸۲
	(۰٫۰۱۵۵)	–	(۰٫۰۰۳۶)	–	–	–	–
$Poisson$	۰٫۱۰۸۵	–	–	۲٫۹۱۲۸	۱۲۹٫۴۳۱۸	۲۶۰٫۸۶۳۶	۲۶۳٫۴۶۸
	(۰٫۰۱۶۴)	–	–	–	–	–	–
DR	−۰٫۰۰۱۵	–	−۰٫۰۰۱۴	–	۱۷۰٫۴۶۳۴	۳۴۲٫۹۲۶۸	۳۴۵٫۵۳۲۰
	(۰٫۰۰۲۲)	–	(۰٫۰۰۰۲)	–	–	–	–
$Poisson$	۰٫۰۱۳۶	–	–	۲٫۹۸۸۹	۱۷۰٫۵۳۹۱	۳۴۳٫۰۷۸۲	۳۴۵٫۶۸۳۴
	(۰٫۰۰۲۸)	–	–	–	–	–	–
DW	۰٫۱۶۴۰	۰٫۱۸۰۴	۰٫۵۶۴۴	–	۱۲۸٫۲۳۱۳	۲۶۰٫۴۶۲۶	۲۶۵٫۶۷۲۹
	(۰٫۰۷۰۱)	(۰٫۰۷۲۴)	(۱٫۰۵۷۲)	–	–	–	–
$Poisson$	۰٫۲۴۵۱	–	–	۲٫۸۹۷۱	۱۲۹٫۲۳۱۵	۲۶۰٫۴۶۳۰	۲۶۳٫۰۶۸۲
	(۰٫۱۲۴۸)	–	–	–	–	–	–
$n=۵۰۰$							
GEO	−۰٫۰۰۳۰	–	−۰٫۰۰۰۴	–	۶۵۱٫۳۲۴۲	۱۳۰۴٫۶۴۸	۱۳۰۸۸۶۳
	(۰٫۰۰۶۳)	–	(۰٫۰۰۱۰)	–	–	–	–
$Poisson$	۰٫۳۶۰۰	–	–	۲٫۷۴۰۲	۶۵۵٫۱۳۵۷	۱۳۱۲٫۲۷۱	۱۳۱۶٫۴۸۶
	(۰٫۱۳۰۴)	–	–	–	–	–	–
DR	−۰٫۰۰۰۷	–	−۰٫۰۰۰۴	–	۸۶۰٫۰۲۷۹	۱۷۲۲٫۰۵۶	۱۷۲۶٫۲۷۰
	(۰٫۰۰۰۴)	–	(۰٫۰۰۰۰)	–	–	–	–
$Poisson$	−۰٫۰۱۲۴	–	–	۲٫۸۷۸۵	۸۶۰٫۲۳۵۳	۱۷۲۲٫۴۷۱	۱۷۲۶٫۶۸۵
	(۰٫۰۰۰۷)	–	–	–	–	–	–
DW	۰٫۰۷۹۰	۰٫۱۴۵۴	۰٫۷۶۷	–	۶۵۱٫۱۴۹۶	۱۳۰۶٫۲۹۹	۱۳۱۴٫۷۲۸
	(۰٫۰۳۳۹)	(۰٫۰۲۲۸)	–	–	–	–	–
$Poisson$	۰٫۳۶۰۴	–	–	۲٫۷۶۷۰	۶۵۵٫۳۸۵۱	۱۳۱۲٫۷۷۰	۱۳۱۶٫۹۸۵
	(۰٫۱۳۰۷)	–	–	–	–	–	–
$n=۱۰۰۰$							
GEO	−۰٫۰۰۲۴	–	۰٫۰۰۰۰	–	۱۳۰۵٫۳۲۲	۲۶۱۲٫۶۴۴	۲۶۱۷٫۵۵۲
	(۰٫۰۰۳۹)	–	(۰٫۰۰۰۵)	–	–	–	–
$Poisson$	۰٫۳۶۱۷	–	–	۲٫۷۴۰۱	۱۳۱۳٫۴۳۶	۲۶۲۸٫۸۷۲	۲۶۳۳٫۷۸۰
	(۰٫۱۳۱۲)	–	–	–	–	–	–
DR	۰٫۰۰۰۰	–	−۰٫۰۰۰۲	–	۱۷۲۲٫۵۶۳	۳۴۴۷٫۱۲۶	۳۴۵۲٫۰۳۴
	(۰٫۰۰۰۲)	–	(۰٫۰۰۰۰)	–	–	–	–
$Poisson$	−۰٫۰۱۱۲	–	–	۲٫۸۷۶	۱۷۲۲٫۹۵۱	۳۴۴۷٫۹۰۲	۳۴۵۲٫۸۱۰
	(۰٫۰۰۰۴)	–	–	–	–	–	–
DW	۰٫۰۴۳۹	۰٫۰۸۲۶	۰٫۰۴۷۱	–	۱۳۰۴٫۱۰۲	۲۶۱۲٫۲۰۴	۲۶۲۲٫۰۱۹
	(۰٫۰۲۳۲)	(۰٫۰۳۵۶)	(۰٫۰۱۲۶)	–	–	–	–
$Poisson$	۰٫۳۶۱۹	–	–	۲٫۷۳۲۶	۱۳۱۲٫۵۵	۲۶۲۷٫۱۰۰	۲۶۳۲٫۰۰۸
	(۰٫۱۳۱۳)	–	–	–	–	–	–

ندارند یعنی توزیع‌های پواسون و توزیع‌های گامای گسسته)، توزیعی که بهترین برازش با این مجموعه داده دارد را انتخاب می‌کنیم، سه معیار $-\log(L)$، AIC و BIC را نیز برای مقایسه بین مدل‌ها در نظر می‌گیریم.

مدل رگرسیون را با استفاده از توزیع انتخاب شده به داده‌ها برازش می‌دهیم. برای بررسی کارایی مدل‌های برازش شده، علاوه بر استفاده از معیارهای قبلی، از معیارهای [۷]خطای مطلق (ABE) و آماره کای–دو پیرسون χ^2 نیز استفاده می‌کنیم.

$$ABE = \sum_{i=1}^{c} |o_i - e_i|,$$
$$\chi^2 = \sum_{i=1}^{c} \frac{(o_i - e_i)^2}{e_i},$$

که در آن، o_i فراوانی مشاهده شده، e_i فراوانی مورد انتظار رده iام، و c تعداد کل رده‌ها است.

۱.۸ مجموعه داده زیست‌شیمی‌دانان

این مجموعه داده از مقاله لانگ [۲۵]، که به بررسی دانشجویان دکترای زیست‌شیمی می‌پردازد استخراج شده است. متغیر پاسخ تعداد فرزندان ۵ ساله یا کوچکتر است که آمار توصیفی این متغیر در جدول ۷ قابل مشاهده است. این مجموعه داده در بسته "*pscl*" نرم‌افزار R نسخه ۴–۳–۰ موجود است.

به نظر می‌رسد داده‌ها در صفر آماسیده هستند، بنابراین توزیع‌های آماسیده در صفر را برازش می‌دهیم. نتایج جدول ۵، تایید می‌کند که توزیع‌های DMW $DRMW$، EP GOM و $Chen$ آماسیده در صفر بهترین مدل‌ها هستند. ما از بین این مدل‌ها، مدل $DRMW$ آماسیده در صفر را برای برازش مدل رگرسیونی انتخاب می‌کنیم. برای برازش مدل رگرسیون، از دو متغیر همراه phd (اعتبار گروه محل تحصیل دکتری) و mar (وضعیت تأهل دانشجو، با سطوح مجرد و متأهل) استفاده می‌کنیم (در ضمن تابع پیوند را فقط روی پارامتر وزن لحاظ می‌کنیم). جدول ۶، MLE پارامترهای مدل، خطای استاندارد برآوردها و آزمون ضرورت برای هر یک از ضرایب رگرسیون را نشان می‌دهد (در این مورد از $GLRT$ استفاده شده است). نتایج نشان می‌دهد که حضور تمام متغیرهای همراه در مدل ضروری است. همچنین، با مقایسه دو معیار AIC و BIC در جداول ۵ و ۶ بهبود مدل بامتغیر همراه نسبت به مدل بدون متغیر همراه نشان داده شده است. جدول ۷ نیز برازش مناسب این مدل رگرسیون را با این مجموعه داده تأیید می‌کند.

همچنین با انجام $GLRT$ برای مقایسه بین برازش توزیع آماسیده و توزیع غیر آماسیده، (توزیع $DRMW$ آماسیده در صفر و توزیع $DRMW$)، آماره آزمون ۱۷۰٫۵۲۲۶ با سطح معنی داری کمتر از ۰٫۰۰۱ بدست آمد که نشان دهنده تفاوت معنی‌دار استفاده از مدل آماسیده و غیر آماسیده است. بنابراین، قطعاً باید توزیع آماسیده را انتخاب کنیم.

همچنین از شکل ۱ می‌توان برای بررسی دقت مدل برازش‌شده استفاده کرد. شکل ۱، نمودار سمت چپ بالایی پراکندگی RQR مدل را برای تمام مشاهدات نشان می‌دهد. این نمودار توزیع باقیمانده‌های تصادفی حول صفر را نشان می‌دهد. نمودار بالایی سمت راست هیستوگرام RQR و نمودار پایینی نمودار QQ را نشان می‌دهد. تمام این شکل‌ها مدل پیشنهادی (مدل رگرسیون $DRMW$ آماسیده در صفر) را پشتیبانی می‌کنند.

۹ نتیجه‌گیری و ملاحظات

در این تحقیق، سعی شده از خانواده توزیع‌های تلسکوپی (که ناشناخته و کمتر استفاده شده‌اند) به عنوان توزیع پایه در توزیع‌های آماسیده استفاده شود. مهم‌ترین ویژگی این خانواده آن است که شامل تعداد زیادی از توزیع‌های آماری متداول می‌شود. همچنین این خانواده از توزیع‌ها شامل خانواده مهمی از توزیع‌های طول عمر گسسته مثل توزیع‌های وایبول گسسته است. بنابراین، این خانواده می‌تواند انتخاب خوبی برای توزیع پایه در مدل‌های آماسیده باشد.

برای ادامه این تحقیق، می‌توان از تعمیم مدل پیشنهادی به حالت چندمتغیره استفاده کرد. همچنین امکان برآورد پارامترهای مدل با سایر روش‌های استنباطی مانند بیزی و بوت‌استرپ وجود دارد. علاوه بر این، با توجه به توانایی این مدل، می‌توان از آن در روش‌های درخت تصمیم، جنگل‌های تصادفی، کنترل کیفیت و غیره استفاده کرد. همچنین به دلیل طولانی نشدن مقاله ما فقط داده‌های گسسته را مورد بررسی قرار دادیم. همانگونه که بیان شد این خانواده از توزیع‌ها با بسیاری از توزیع‌های پیوسته نیز در ارتباط است. پس می‌توان مدل‌های پیوسته آماسیده را نیز برای آینده مورد مطالعه قرار داد.

[7] Absolute Error

جدول ۵: انتخاب بهترین مدل در خانواده توزیع تلسکوپی

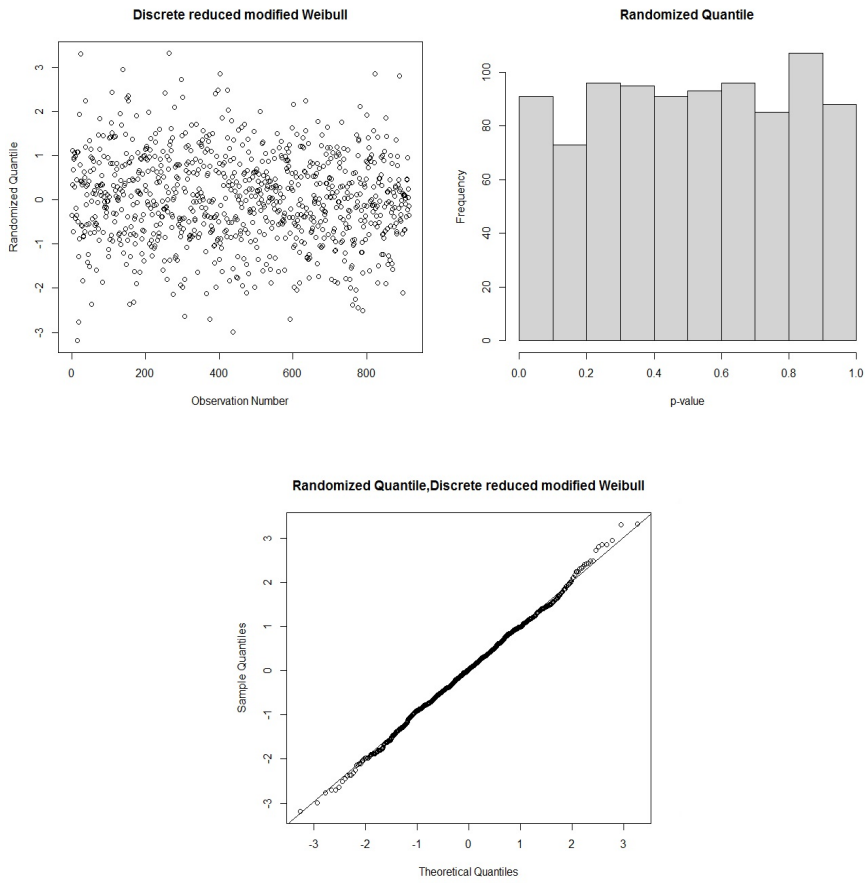
توزیع	$-\log(L)$	AIC	BIC
E	۸۶۹٫۷۸۳۴	۱۷۴۳٫۵۶۷	۱۷۵۳٫۲۰۵
GEO	۸۶۸٫۶۵۷۸	۱۷۴۱٫۳۱۶	۱۷۵۰٫۹۵۳
DW	۸۳۷٫۸۱۶۳	۱۷۰۱٫۶۳۳	۱۷۱۶٫۰۸۹
DRA	۸۵۰٫۲۴۹۰	۱۷۰۴٫۴۹۸	۱۷۱۴٫۱۳۶
LE	۸۶۰٫۶۴۰۶	۱۷۲۹٫۲۸۱	۱۷۴۸٫۵۵۷
MW	۸۴۸٫۹۲۳۵	۱۷۰۵٫۸۴۷	۱۷۲۵٫۱۲۳
DMW	۸۴۷٫۴۹۳۶	۱۷۰۲٫۹۸۷	۱۷۲۲٫۲۶۳
DRMW	۸۴۷٫۳۸۰۸	۱۷۰۲٫۷۶۲	۱۷۲۲٫۰۳۷
Chen	۸۴۷٫۵۲۹۸	۱۷۰۱٫۰۶۰	۱۷۱۵٫۵۱۶
EE	۸۶۰٫۱۱۵۵	۱۷۲۶٫۲۳۱	۱۷۴۰٫۶۸۸
EP	۸۴۷٫۷۶۳۱	۱۷۰۱٫۵۲۶	۱۷۱۵٫۹۸۳
GOM	۸۴۷٫۷۷۱۳	۱۷۰۱٫۵۴۳	۱۷۱۵٫۹۹۹
GPW	۸۵۷٫۱۸۸۴	۱۷۲۰٫۳۷۷	۱۷۳۴٫۸۳۴
LFR	۸۵۳٫۱۷۰۹	۱۷۱۲٫۳۴۲	۱۷۲۶٫۷۹۹
LG	۸۸۰٫۱۷۸۳	۱۷۶۶٫۳۵۷	۱۷۸۰٫۸۱۳
WEE	۸۵۳٫۰۵۶۴	۱۷۱۲٫۱۱۳	۱۷۲۶٫۵۷۰
BX	۸۵۵٫۱۰۰۰	۱۷۱۶٫۲۰۰	۱۷۳۰٫۶۵۷
Poisson	۸۵۶٫۴۲۳۱	۱۷۱۶٫۸۴۶	۱۷۲۶٫۴۸۴
DGamma	۸۴۸٫۳۳۴۹	۱۷۰۲٫۶۷۰	۱۷۱۷٫۱۲۷

جدول ۶: مدل رگرسیونی مبتنی بر توزیع *DRMW* آماسیده در صفر برای مجموعه داده زیست‌شیمی‌دانان

پارامتر	*MLE*	*SE*	سطح معنی‌داری
β_0	$-۱۸٫۰۱۹۳$	۰٫۰۱۲۲	$< ۰٫۰۰۱$
β_1	$-۱٫۸۱۴۴$	۰٫۰۱۴۵	$< ۰٫۰۰۱$
β_2	$-۰٫۰۲۱۱$	۰٫۰۱۶۰	$< ۰٫۰۰۱$
α	۲۲٫۲۳۶۶	۰٫۹۸۳۹	–
γ	۱٫۸۵۲۴	۰٫۰۳۶۳	–
q	۰٫۹۸۵۵	۰٫۰۰۰۶	–
$-\log(L)$	۶۷۶٫۸۵۸۱۶	–	–
AIC	۱۳۶۷٫۷۱۶	–	–
BIC	۱۴۰۱٫۴۴۹	–	–

جدول ۷: تعداد کودکان ۵ سال یا کمتر و مقادیر مورد انتظار مربوطه تحت مدل‌های برازش شده در جدول ۶

مقدار	مشاهده شده	برآورد
۰	۵۹۹	۵۹۹٫۵۵
۱	۱۹۵	۱۹۳٫۳۹
۲	۱۰۵	۱۰۶٫۶۵
$۳<$	۱۶	۱۵٫۲۴
ABE		۷٫۵۵۳۷
χ^2		۰٫۱۱۱۳



شکل ۱: نمودارهای پراکندگی، هیستوگرام و QQ برای RQR های مجموعه داده‌های بیوشیمی‌دانان

ضمیمه

$$\frac{\partial^2 \ell(w,q,m_\theta;\mathbf{y})}{\partial w^2} = \sum_{j=1}^{n} \left\{ -I(y_j) \left(\frac{1 - g(y_j;q,m_\theta)}{\Lambda_j} \right)^2 - \frac{I(y_j > k)}{\eta^2(w)} \right\},$$

$$\begin{aligned} \frac{\partial^2 \ell(w,q,m_\theta;\mathbf{y})}{\partial \theta^2} = & \sum_{j=1}^{n} \left\{ \left[I(y_j) \times \frac{\eta(w)\partial^2 g(y_j;q,m_\theta)/\partial\theta^2 \Lambda_j - (\eta(w)\partial g(y_j;q,m_\theta)/\partial\theta)^2}{\Lambda_j^2} \right] \right. \\ & + \left. I(y_j > 0) \times \frac{\partial^2 g(y_j;q,m_\theta)/\partial\theta^2 g(y_j;q,m_\theta) - (\partial g(y_j;q,m_\theta)/\partial\theta)^2}{g^2(y_j;q,m_\theta)} \right\}, \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 \ell(w,q,m_\theta;\mathbf{y})}{\partial q^2} = & \sum_{j=1}^{n} \left\{ \left[I(y_j) \times \frac{\eta(w)\partial^2 g(y_j;q,m_\theta)/\partial q^2 \Lambda_j - (\eta(w)\partial g(y_j;q,m_\theta)/\partial q)^2}{\Lambda_j^2} \right] \right. \\ & + \left. I(y > k) \times \frac{\partial^2 g(y_j;q,m_\theta)/\partial q^2 g(y_j;q,m_\theta) - (\partial g(y_j;q,m_\theta)/\partial q)^2}{g^2(y_j;q,m_\theta)} \right\}, \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 \ell(w,q,m_\theta;\mathbf{y})}{\partial w \partial \theta} = & \sum_{j=1}^{n} \left\{ \left[-I(y_j) \times \frac{(\partial g(y_j;q,m_\theta)/\partial\theta)[\Lambda_j - \eta(w) g(y_j;q,m_\theta)]}{\Lambda_j^2} \right] \right. \\ & - \left. I(y_j) \frac{(\partial g(y_j;q,m_\theta)/\partial\theta)[\Lambda_j + \eta(w)(1 - g(y_j;q,m_\theta))]}{\Lambda_j^2} \right\}, \end{aligned}$$

$$\frac{\partial^{۲}\ell(w,q,m_\theta;\mathbf{y})}{\partial w\partial q} = \sum_{j=۱}^{n}\left\{\left[-I(y_j)\times\frac{(\partial g(y_j;q,m_\theta)/\partial q)[\Lambda_j-\eta(w)g(y_j;q,m_\theta)]}{\Lambda_j^{۲}}\right]\right.$$
$$\left. - \; I(y_j)\frac{(\partial g(y_j;q,m_\theta)/\partial q)[\Lambda_j+\eta(w)(۱-g(y_j;q,m_\theta))]}{\Lambda_j^{۲}}\right\},$$

$$\frac{\partial^{۲}\ell(w,q,m_\theta;\mathbf{y})}{\partial q\partial\theta} = \sum_{j=۱}^{n}\left\{\left[-I(y_j)\times\frac{(\partial^{۲}g(y_j;q,m_\theta)/\partial\theta\partial q)\Lambda_j\eta(w)-(\partial g(y_j;q,m_\theta)/\partial\theta)(\partial g(y_j;q,m_\theta)/\partial q)\eta^{۲}(w)}{\Lambda_j^{۲}}\right]\right.$$
$$\left. + \; I(y_j>۰)\frac{(\partial^{۲}g(y_j;q,m_\theta)/\partial\theta\partial q)-(\partial g(y_j;q,m_\theta)/\partial\theta)(\partial g(y_j;q,m_\theta)/\partial q)}{g^{۲}(y_j;q,m_\theta)}\right\},$$

که در آن، $\Lambda_j = w+\eta(w)g(y_j;q,m_\theta),\ j=۱,\cdots,n$.

مراجع

[1] Agarwal, D. K., Gelfand, A. E., & Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, **9**, 341–355.

[2] Agresti, A., & Kateri, M. (2025). Categorical data analysis. In *International Encyclopedia of Statistical Science* (pp. 408–411). Springer, Berlin.

[3] Almalki, S. J., & Yuan, J. (2013). A new modified Weibull distribution. *Reliability Engineering and System Safety*, **111**, 164–170.

[4] Almalki, S. J., & Nadarajah, S. (2014). A new discrete modified Weibull distribution. *IEEE Transactions on Reliability*, **63**, 68–80.

[5] Bain, L. J. (1974). Analysis for the linear failure-rate life-testing distribution. *Technometrics*, **16**(4), 551–559.

[6] Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Cambridge University Press.

[7] Chen, Z. (1999). Statistical inference about the shape parameter of the exponential power distribution. *Statistical Papers*, **40**(4), 459–468.

[8] Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics and Probability Letters*, **49**, 155–161.

[9] Chen, X. D. (2009). Bayesian analysis of semiparametric mixed-effects models for zero-inflated count data. *Communications in Statistics - Theory and Methods*, **38**, 1815–1833.

[10] Dagne, G. A. (2010). Bayesian semiparametric zero-inflated Poisson model for longitudinal count data. *Mathematical Biosciences*, **224**, 126–130.

[11] Edwin, T. K. (2014). *Power Series Distributions and Zero-Inflated Models*. Unpublished Thesis, University of Nairobi.

[12] Feng, C., Sadeghpour, A., & Li, L. (2017). Randomized quantile residuals: an omnibus model diagnostic tool with unified reference distribution. *Journal of Computational and Graphical Statistics*.

[13] Ghosh, S. K., Mukhopadhyay, P., & Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, **136**, 1360–1375.

[14] Gupta, R. D., & Kundu, D. (2009). A new class of weighted exponential distributions. *Statistics*, **43**(6), 621–634.

[15] Jansakul, N., & Hinde, J. P. (2002). Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis*, **40**, 75–96.

[16] Jung, B. C., Jhun, M., & Lee, J. W. (2005). Bootstrap tests for overdispersion in a zero-inflated Poisson regression model. *Biometrics*, **61**, 626–629.

[17] Katti, S. K., & Rao, A. V. (1970). The log-zero-Poisson distribution. *Biometrics*, **26**, 801–813.

[18] Khoshgoftaar, T. M., Gao, K., & Szabo, R. M. (2005). Comparing software fault predictions of pure and zero-inflated Poisson regression models. *International Journal of Systems Science*, **36**, 705–715.

[19] Klugman, S., Panjer, H., & Willmot, G. (2004). *Loss Models: From Data to Decisions* (2nd ed.). Wiley, New York.

[20] Kundu, D., & Raqab, M. Z. (2005). Generalized Rayleigh distribution: different methods of estimation. *Computational Statistics and Data Analysis*, **49**, 187–200.

[21] Lai, C. D., Xie, M., & Murthy, D. N. P. (2003). A modified Weibull distribution. *IEEE Transactions on Reliability*, **52**(1), 33–37.

[22] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

[23] Lim, H. K., Li, W. K., & Yu, P. L. (2014). Zero-inflated Poisson regression mixture model. *Computational Statistics and Data Analysis*, **71**, 151–158.

[24] Long, D. L., Preisser, J. S., Herring, A. H., & Golin, C. E. (2014). Zero-inflated Poisson regression with application to defects in manufacturing. *Statistics in Medicine*, **33**, 5151–5165.

[25] Long, J. S. (1990). The origins of sex differences in science. *Social Forces*, **68**(3), 1297–1316.

[26] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **44**(2), 226–233.

[27] Marshall, A. W., & Olkin, I. (2007). *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families*. Springer, New York.

[28] Murthy, D. N. P., Xie, M., & Jiang, R. (2003). *Weibull Models*. Wiley, New York.

[29] Musio, M., Sauleau, E. A., & Buemi, A. (2010). Bayesian semiparametric ZIP models with space-time interactions: an application to cancer registry data. *Mathematical Medicine and Biology*, **27**, 181–194.

[30] Nakagawa, T., & Osaki, S. (2009). The discrete Weibull distribution. *IEEE Transactions on Reliability*, **24**(5), 300–301.

[31] Neelon, B., & Chung, D. J. (2017). The LZIP: a Bayesian latent factor model for correlated zero-inflated counts. *Biometrics*, **73**, 185–196.

[32] Nikulin, M., & Haghighi, F. (2006). A Chi-squared test for the generalized power Weibull family for the head-and-neck cancer censored data. *Journal of Mathematical Sciences*, **133**(3), 1333–1341.

[33] Nooghabi, M. S., Borzadaran, G. R. M., & Roknabadi, A. H. R. (2011). Discrete modified Weibull distribution. *Metron*, **69**, 207–222.

[34] Pham, H. (2002). A Vtub-shaped hazard rate function with applications to system safety. *International Journal of Reliability and Applications*, **3**(1), 1–16.

[35] Ridout, M., Hinde, J., & Demétrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**, 219–223.

[36] Roknabadi, A. R., Borzadaran, G. R. M., & Khorashadizadeh, M. (2009). Some aspects of discrete hazard rate function in Telescopic families. *Journal of Statistical Research*, **43**(2), 1–15.

[37] Roy, D. (2004). Discrete Rayleigh distribution. *IEEE Transactions on Reliability*, **53**(2), 255–260.

[38] Saboori, H., & Doostparast, M. (2022). Flexible multivariate zero to k inflated power series regression model with applications. *Stat*, **11**(1), e473.

[39] Sakia, R. M. (2018). Application of the power series probability distributions for the analysis of zero-inflated insect count data. *Open Access Library Journal*, **5**, e4735.

[40] Sarhan, A. M., & Zaindin, M. (2009). Modified Weibull distribution. *Applied Sciences*, **11**, 123–136.

[41] Vandenbroek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**, 738–743.

[42] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, **11**, 95–103.

[43] Xiao, X., Tang, Y., Xu, A., & Wang, G. (2020). Bayesian inference for zero-and-one-inflated geometric distribution regression model using Pólya-Gamma latent variables. *Communications in Statistics - Theory and Methods*, **49**, 3730–3743.

[44] Xie, M., & Lai, C. D. (1996). Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. *Reliability Engineering and System Safety*, **52**(1), 87–93.

[45] Zhu, H., Luo, S., & DeSantis, S. M. (2017). Zero-inflated count models for longitudinal measurements with heterogeneous random effects. *Statistical Methods in Medical Research*, **26**, 1774–1786.

Zero-Inflated Telescopic Models and Applications

H. Saboori[1]* and M. Doostparast[2]

[1] Department of Statistics, University of Zabol, Zabol, Iran

[2] Department of Statistics, Ferdowsi University of Mashhad, Iran

Abstract:

Discrete inflated data are widely used in practice. One of the most essential approaches for modeling such data involves the use of models based on inflated distributions. Since the choice of the baseline distribution plays a fundamental role in defining a family of inflated distributions, the efficiency and fitting capability of the model are directly related to the baseline distribution. In most research concerning inflated data, models based on the Poisson distribution are employed. The Poisson distribution is a very powerful distribution; however, it has one characteristic that becomes its Achilles' heel in application. This characteristic is the equality of the mean and variance—a condition that rarely holds in real-world data. Therefore, it is necessary to seek an alternative to the Poisson distribution, and what better alternative than a broad family of distributions with greater flexibility? One such candidate is the family of Telescopic distributions.

In this research, the family of Telescopic distributions is used as the baseline distribution to define a general class of discrete inflated models. This family includes distributions that have been less frequently used as baseline distributions (such as the discrete Weibull distribution). Moreover, these distributions are also used as lifetime distributions. Additionally, due to their connection with an important family of continuous distributions, they encompass a wide spectrum of other statistical distributions. The breadth of this family is a significant advantage. Thus, if the Telescopic family is used as the baseline distribution for defining inflated models, there will be numerous new models available for fitting any discrete inflated (and even non-inflated) dataset. It is worth noting that this research also briefly addresses a regression model based on the mixing parameter.

Keywords: Inflated data, Inflated distributions, Generalized linear models, Family of telescoping distributions.

*Corresponding author: saboori_hadi@yahoo.com