

انتخاب متغیر با رویکرد جدید در آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری با توزیع پواسون

ملیحه حیدری و فرزاد اسکندری

گروه آمار، دانشگاه علامه طباطبایی

تاریخ دریافت: ۱۳۹۵/۲/۲۲ تاریخ آخرین بازنگری: ۱۳۹۶/۶/۲

چکیده: در این مقاله به بحث انتخاب متغیر با رویکردی جدید در آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری پرداخته می‌شود، به گونه‌ای که داده‌ها از توزیع پواسون تبعیت می‌کنند. اما دو عامل بیش‌پراکندگی و صفرهای بیش از حد به دلیل استفاده از توزیع پواسون می‌تواند تاثیر زیادی بر انتخاب متغیر و برآورد پارامترها داشته باشند. در واقع برآورد پارامترها در بخش پارامتری مدل رگرسیونی نیم‌پارامتری با استفاده از رویکرد درست‌نمایی تاوانیده انجام می‌پذیرد و در بخش ناپارامتری پس از تقریب موضعی تابع ناپارامتری با استفاده از بسط تیلور، محاسبات در حضور برآورد ضرایب پارامتری انجام می‌گیرد. استفاده از رویکرد جدید در این مقاله باعث شده است تا موانع در انتخاب درست متغیرها برطرف گردد. در این مقاله علاوه بر ارائه تئوری‌های مربوطه، در بخش شبیه‌سازی داده‌ها نیز دو موضوع بیش‌پراکندگی و صفرهای بیش از حد مورد توجه قرار می‌گیرد و استفاده از روش EM در برآورد پارامترها منجر به افزایش دقت در نتیجه شده است.

واژه‌های کلیدی: الگوریتم EM، بیش‌پراکندگی، صفرهای بیش از حد، رگرسیون نیم‌پارامتری، مدل آمیزه‌ای متناهی.

۱ مقدمه

مدت زمانی طولانی است که مسئله تعیین بهترین زیرمجموعه از متغیرها در یک مدل آماری، موضوع مورد علاقه و کاربردی آمارشناسان شده است. بر همین مبنا و بر اساس یک مدل رگرسیونی ساده، رگرسیون کم‌ترین توان‌های دوم عادی^۱ (OLS) مجموع توان دوم مانده‌ها^۲ $RSS = (y - x\beta)^T(y - x\beta)$ را مینیمم می‌کند و براوردی نااریب به فرم $\hat{\beta}_{OLS} = (x^T x)^{-1} x^T y$ حاصل می‌کند. در حالتی که ماتریس طرح X پرتبه نباشد، هم‌خطی بین اجزای ماتریس رخ می‌دهد که نشان از همبستگی میان متغیرهای مستقل است و هرچه میزان آن افزایش یابد، واریانس براورد ضرایب نیز افزایش خواهد یافت، در نتیجه براوردگر (OLS) یکتا نخواهد بود. از طرف دیگر با استفاده از براورد $\hat{\beta}_{OLS}$ ، میانگین توان دوم خطاها در زمان پیش‌بینی افزایش می‌یابد. بنابراین از نظر محاسباتی این شیوه‌های سنتی مقرون به صرفه نیستند و نتیجه‌ای رضایت‌بخش در پی نخواهند داشت.

بنابراین محققان برای حل مشکل، شیوه انتخاب متغیر تاوانیده را مطرح کردند که با ایجاد محدودیت روی مجموع توان دوم مانده‌ها، پارامترها را براورد می‌کند. از جمله این محققان تیشیرانی (۱۹۹۶) است که تابع تاوانی با نام کمترین انقباض مطلق و عملگر گزینش^۳ (LASSO) را معرفی کرد. سپس فن و لی (۲۰۰۱) تابع تاوان انحراف مطلق به طور هموار بریده‌شده^۴ (SCAD) را معرفی کردند، به طوری که در حضور این تابع‌های تاوان عمل انتخاب متغیر و براورد پارامترها به طور همزمان انجام می‌پذیرد. امروزه نیاز به مدل‌های آماری انعطاف‌پذیر مانند مدل‌های رگرسیونی نیم‌پارامتری به دلیل حجم وسیعی از داده‌های پزشکی، مهندسی و محیطی افزایش یافته است. چون با استفاده از این مدل‌ها می‌توان به تحلیل داده‌های پیچیده در حال توسعه پرداخت، به خصوص زمانی که داده‌ها برخاسته از یک جامعه ناهمگن باشند. مک‌لاکلان و پیل (۲۰۰۰) مرور جامعی بر روی این‌گونه مدل‌ها و چگونگی ساختار آنها داشته‌اند. خلیلی و چن (۲۰۰۷) بحث انتخاب متغیر با رویکرد تاوانیده را بر روی آمیزه‌ای متناهی از یک مدل رگرسیونی ساده مورد مطالعه قرار دادند.

هاردل و همکاران (۲۰۰۰)، راپرت و همکاران (۲۰۰۳) و یاچو (۲۰۰۳) مدل‌های رگرسیونی نیم‌پارامتری را به همراه شیوه‌های استنباط آن‌ها مورد بررسی قرار داده‌اند. لی و لیانگ (۲۰۰۸) مسئله

^۱Ordinary Least Squares

^۲Residual Sum of Squares

^۳Least Absolute Shrinkage and Selection Operator

^۴Smoothly Clipped Absolute Deviation

انتخاب متغیر با رویکرد تاوانیده در زمینه مدل‌های خطی جزئی ضریب-متغیر تعمیم‌یافته^۵ (GVCPLM) را مورد بررسی قرار دادند. انتخاب متغیر با رویکرد تاوانیده در ارتباط با این نوع مدل‌ها چالش‌برانگیز است، چون شامل انتخاب متغیرهای معنادار برای هر دو بخش پارامتری و ناپارامتری است. هانتز و یانگ (۲۰۱۲) آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری را در نظر گرفتند و در مورد خطاهای هر یک از زیرمدل‌ها فرض کردند که خطاها مستقل و هم‌توزیع باشند، هم‌چنین بخش پارامتری را به سه حالت مختلف بسط دادند و با استفاده از الگوریتم شبه EM به برآورد پارامترها پرداختند. اسکندری و اورمز (۲۰۱۵) بحث انتخاب متغیر در خانواده توزیع‌های نمایی را با استفاده از شیوه‌های تاوانیده در ترکیبی از مدل تعمیم‌یافته نیم‌پارامتری لی و لیانگ (۲۰۰۸) و آمیزه‌ای متناهی از مدل‌های رگرسیونی خلیلی و جن (۲۰۰۷) مطرح کردند که تا آن زمان، شیوه‌های تاوانیده ارائه‌شده در مورد ترکیب دو مدل مطرح‌شده به طور هم‌زمان بررسی نشده بودند. اورمز و اسکندری (۲۰۱۶) انتخاب متغیر در آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری تعمیم‌یافته با استفاده از برآورد تاوانیده را مطرح کردند، آنها مدل ربط را در حالت نیم‌پارامتری بسط دادند و تابع ناپارامتری را چندبعدی در نظر گرفتند.

هدف این مقاله انتخاب متغیر در آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری بر مبنای یک رویکرد جدید و در شرایطی است که داده‌ها از توزیع پواسون تبعیت می‌کنند. در واقع از بین خانواده توزیع‌های نمایی، حالت خاص توزیع پواسون در نظر گرفته می‌شود، به این دلیل که ممکن است مشکل بیش‌پراکندگی^۶ و صفرهای بیش از حد^۷ رخ دهد و محاسبات را از روند طبیعی خود خارج کند. در چنین شرایطی استفاده از توزیع دوجمله‌ای منفی به جای استفاده از توزیع پواسون پیشنهاد می‌شود. ما در این مقاله وارد بحث تئوری آن نمی‌شویم بلکه این راهکار را به عنوان یک رویکرد جدید در برآورد پارامترهای مدل مطرح شده به کار می‌گیریم و نشان می‌دهیم که با جایگذاری توزیع دوجمله‌ای منفی مشکل بیش‌پراکندگی و صفرهای بیش از حد برطرف می‌گردد، به طوری که در برآورد صحیح پارامترها مشکلی ایجاد نمی‌شود.

صفرهای بیش از حد و بیش‌پراکندگی پدیده‌های رایجی هستند که در استفاده از مدل‌های رگرسیونی پواسون ساده به منظور مدل‌بندی داده‌های شمارشی محدودیت ایجاد می‌کنند. این مشکل زمانی رخ می‌دهد که نمونه‌های متغیر پاسخ، حاصل از جامعه‌ای با چندین زیرجامعه باشند، در نتیجه ناهمگنی غیر قابل مشاهده در داده‌ها ایجاد می‌شود. به منظور کنترل بیش‌پراکندگی استفاده از یک مدل رگرسیونی دوجمله‌ای منفی پیشنهاد می‌شود، چون در این حالت واریانس از میانگین بزرگتر است.

^۵Generalized Varying-Coefficient Partially Linear Model

^۶Overdispersion

^۷Excess zeros

یک شیوه رایج به منظور مدل‌بندی صفرهای بیش از حد، استفاده از مدل پواسونی با انباشتگی صفر بیش از حد^۸ (ZIP) است که توسط لامبرت (۱۹۹۲) مورد مطالعه قرار گرفته است. لیم و همکاران (۲۰۱۴) اظهار داشتند که توزیع پواسون متورم صفر، آمیزه‌ای از یک توزیع پواسون و یک توزیع تباهیده در نقطه صفر است. علاوه بر این اگر مجدداً بیش‌پراکنشی بعد از مدل‌بندی صفرهای اضافی وجود داشته باشد، راه حل مناسب ارائه‌شده توسط لیم و همکاران (۲۰۱۴) استفاده از یک مدل دوجمله‌ای منفی با انباشتگی صفر بیش از حد^۹ (ZINB) است.

در بخش ۲ مدل‌های آمیزه‌ای متناهی با تعریف صورت پارامتری آنها معرفی می‌شوند. در بخش ۳ حالت رگرسیونی مدل‌های آمیزه‌ای متناهی به شکل نیم‌پارامتری و به نام آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری تشریح می‌شود. بخش ۴ شامل مطالعه شبیه‌سازی به منظور بررسی عملکرد الگوریتم EM در برآورد ضرایب مورد نظر است.

۲ مدل‌های آمیزه‌ای متناهی

توزیع‌های آمیزه‌ای متناهی رهیافتی آماری ایجاد می‌کنند تا بر مبنای آن بتوان به مدل‌بندی گسترده وسیعی از پدیده‌های تصادفی پرداخت. به دلیل انعطاف‌پذیری این‌گونه توزیع‌ها برای مدل‌بندی، مدل‌های آمیزه‌ای متناهی از نظر کاربردی و نظری توجه بسیاری را در طی سال‌های اخیر به خود جلب کرده‌اند.

تعریف ۱: فرض کنید Y_1, \dots, Y_n نمونه‌ای تصادفی به اندازه n باشد، به طوری که Y_j بردار تصادفی p بعدی با تابع چگالی احتمال $f(y_j)$ بر روی فضای \mathbb{R}^p است. اگر $f_i(y_j)$ تابع چگالی زامین مشخصه مربوط به i امین مولفه باشد، آنگاه فرم پارامتری مدل‌های آمیزه‌ای به صورت

$$f(y_j; \Psi) = \sum_{i=1}^g \pi_i f_i(y_j; \theta_i) \quad j = 1, \dots, P$$

است که در آن $\Psi = (\pi_1, \dots, \pi_{g-1}, \xi^T)^T$ شامل پارامترهای مدل، شامل $(\theta_1, \dots, \theta_g)$ و نسبت‌های آمیختگی^{۱۰} با وزن‌های π_i است، به طوری که مجموعشان به ازای تمام مولفه‌ها برابر با ۱ است.

^۸Zero-Inflated Poisson

^۹Zero-Inflated Negative Binomial

^{۱۰}Mixing proportion

۳ آمیزه‌های متناهی از مدل‌های رگرسیونی نیم‌پارامتری

بر مبنای مدل معرفی شده توسط اورمز و اسکندری (۲۰۱۶) فرض کنید متغیر پاسخ در ارتباط با متغیر کمکی از مدل رگرسیونی نیم‌پارامتری GVCPLM تبعیت کند. به همین منظور متغیر پاسخ Y با مقادیر ممکن $Y \subset \mathbb{R}$ و برداری از متغیرهای کمکی تاثیرگذار بر روی متغیر پاسخ را به صورت (u, x, z) در نظر بگیرید، که در آن $x = (x_1, \dots, x_q)^T$ ، $z = (z_1, \dots, z_p)^T$ و u یک تک‌متغیر است.

تعریف ۲: فرض کنید $G = \{f(y; \theta, \phi); (\theta, \phi) \in \Theta \times (0, \infty)\}$ خانواده‌ای از تابع‌های چگالی پارامتری Y نسبت به اندازه σ متناهی ν باشد، که در آن $\theta \in \mathbb{R}$ و ϕ یک پارامتر پراکندگی^{۱۱} است. گفته می‌شود (u, x, z, Y) از آمیزه‌های متناهی شامل مدل‌های رگرسیونی نیم‌پارامتری با مرتبه K پیروی می‌کند، هرگاه تابع چگالی شرطی Y به شرط (u, x, z) به صورت

$$f(y; u, x, z, \Psi) = \sum_{k=1}^K \pi_k f(y; \theta_k(u, x, z), \phi_k) \quad k = 1, 2, \dots, K \quad (1)$$

باشد، به گونه‌ای که $\theta_k(u, x, z) = h(x^t \alpha_k(u) + z^t \beta_k)$ با تابع پیوند معلوم $h(\cdot)$ به مدل رگرسیونی نیم‌پارامتری معرفی شده توسط لی و لیانگ (۲۰۰۸) ارتباط پیدا می‌کند و $\alpha(\cdot)$ برداری شامل تابع‌های نامعلوم از ضرایب رگرسیونی هموار است. بردار پارامتری $\Psi = (\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k, \phi, \pi)$ 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 $\pi = (\pi_1, \dots, \pi_{k-1})^T$ است.

مدل‌های رگرسیونی نیم‌پارامتری آمیزه‌ای متناهی^{۱۲} (FMSPR) ارائه شده توسط اورمز و اسکندری (۲۰۱۶) که عضو خانواده نمایی هستند، در تعریف ۲ روشی معمول به منظور مدل‌بندی این‌گونه رابطه‌های ناهمگن نامشهود فراهم می‌کند.

۳،۱ آمیزه‌های متناهی از مدل‌های رگرسیونی نیم‌پارامتری با توزیع پواسون

بر مبنای تعریف ۲، مدل FMSPR که در آن متغیر پاسخ دارای توزیع پواسون با پارامتر λ باشد، همانند (۱) تعریف می‌شود با این تفاوت که به جای $\theta_k(u, x, z)$ ، $\lambda_k(u, x, z)$ در نظر گرفته می‌شود. وقتی

^{۱۱}Dispersion parameter

^{۱۲}Finite Mixture of Semi-Parametric Regression

۶۲ آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری

متغیر تصادفی Y دارای توزیع پواسون با پارامتر λ باشد، پارامتر پراکندگی ϕ_k مقداری برابر ۱ اختیار می‌کند. به دلیل این‌که با تعریف توزیع پواسون به فرم خانواده نمایی (۲)، مقدار پارامتر پراکندگی بر مبنای توزیع پواسون به صورت

$$f_y(y; \theta, \phi) = \exp\left\{\frac{y\theta - B(\theta)}{\phi} + C(y, \phi)\right\} \quad (۲)$$

مشخص می‌شود که در آن $B(\cdot)$ و $C(\cdot, \cdot)$ تابع‌های معلوم هستند و بازه مقادیر y به θ یا ϕ بستگی ندارد. در این فرمول‌بندی، θ پارامتر کانونی و ϕ پارامتر پراکندگی است. بر مبنای تعریف به فرم خانواده نمایی تابع چگالی $f(y_i; \lambda_k(u_i, x_i, z_i), 1)$ به ازای $i = 1, \dots, n$ بر اساس تابع چگالی توزیع پواسون به صورت زیر تعریف می‌شود:

$$f(y_i; \lambda_k(u_i, x_i, z_i), 1) = \exp\{y_i \log \lambda_k - \lambda_k - \log y_i!\}$$

۳,۲ انتخاب متغیر در آمیزه‌ای متناهی از مدل‌های رگرسیونی نیم‌پارامتری با توزیع پواسون

مسئله انتخاب متغیر با رویکرد درست‌نمایی تاوانیده و با در نظر گرفتن توزیع پواسون برای متغیر پاسخ در سه مرحله اصلی بیان می‌شود.

مرحله اول : لگاریتم تابع درست‌نمایی به شرط پارامتر Ψ بر مبنای مدل FMSPR به صورت

$$\ell_n(\Psi) = \sum_{i=1}^n \log\left\{\sum_{k=1}^K \pi_k f(y_i; \lambda_k(u_i, x_i, z_i), 1)\right\}$$

تعریف می‌شود در نتیجه برای برآورد ضرایب مجهول بر مبنای داده‌های کامل در حضور متغیرهای نشانگر فرضی ν_{ik} ، لگاریتم تابع درست‌نمایی کامل به صورت

$$\ell_n^c(\Psi) = \sum_{i=1}^n \sum_{k=1}^K \nu_{ik} \{\log \pi_k + y_i \log \lambda_k - \lambda_k - \log y_i!\}$$

تعریف می‌شود. به دلیل کار کردن با یک مدل رگرسیونی نیم‌پارامتری و عدم اطلاع در مورد تابع ناپارامتری $\alpha(\cdot)$ بر مبنای روش لی و لیانگ (۲۰۰۸)، با استفاده از تقریب موضعی بسط تیلور مرتبه دوم، $\alpha_{k,j}(\nu)$

به ازای ν و در همسایگی متغیری مانند u به صورت

$$\begin{aligned} \alpha_{kj}(\nu) &\approx \alpha_{kj}(u) + \alpha'_{kj}(u)(\nu - u), \quad k = 1, \dots, K, j = 1, \dots, P \\ &\equiv a_{kj} + b_{kj}(\nu - u) \end{aligned} \quad (۳)$$

که در آن تابع $\alpha(\cdot)$ به ازای متغیر u و مولفه k و بعد j ، با متغیر a_{kj} و مشتق مرتبه اول آن با متغیر b_{kj} نمایش داده می شود. بسط تیلور در رابطه (۳) تنها تا مرتبه دوم در نظر گرفته می شود و این در حالی است که هر چه تعداد مرتبه افزایش یابد، دقت نیز افزایش خواهد یافت. بنابراین تابع درستنمایی بر مبنای مولفه ناپارامتری و تابع هسته $k_h(u_i - u) = \frac{1}{h} k\left(\frac{u_i - u}{h}\right)$ ، برای برآورد موضعی و تقریبی β و b, a به صورت

$$\ell_n = \sum_{i=1}^n \log \sum_{k=1}^K \{ \pi_k [\exp\{y_i \log \tilde{\lambda}_k - \tilde{\lambda}_k - \log y_i!\}] k_h(u_i - u) \} \quad (۴)$$

تعریف می شود، که در آن $\tilde{\lambda}_k(u_i, x_i, z_i)$ بر مبنای a و b به شکل

$$\tilde{\lambda}_k(x_i, u_i, z_i) = \exp(x_i^T a_k + x_i^T b_k(u_i - u) + z_i^T \beta_k), \quad i = 1, \dots, n, k = 1, \dots, K \quad (۵)$$

تعریف می شود. پس از مشخص کردن $\tilde{\lambda}_k$ ، مقادیر بهینه (۴) با استفاده از الگوریتم امید ریاضی-ماکسیم سازی^{۱۳} (EM) محاسبه می شود. برای ماکسیم سازی و محاسبه مقدار بهینه، لگاریتم تابع درستنمایی کامل به صورت

$$\ell_n^c(\Psi) = \sum_{i=1}^n \sum_{k=1}^K \nu_{ik} \{ \log \pi_k + (y_i \log \tilde{\lambda}_k - \tilde{\lambda}_k - \log y_i!) + \log k_h(u_i - u) \} \quad (۶)$$

تعریف می شود. در تابع هسته $k_h(u_i - u)$ مندرج در (۶)، h پهنای باند است که با اختیار مقادیر مختلف ما را ملزم به حرکت در فاصله ای معین و در همسایگی u می کند.

گام E: در این گام امید ریاضی شرطی $\ell_n^c(\Psi)$ به شرط ν_{ik} و بر مبنای مشاهدات (u_i, x_i, z_i, y_i)

^{۱۳}Expectation-Maximization

به صورت

$$Q(\Psi; \Psi_{(m)}) = \sum_{i=1}^n \sum_{k=1}^K \omega_{ik}^{(m)} [y_i \log \tilde{\lambda}_k - \tilde{\lambda}_k - \log y_i!] + \sum_{i=1}^n \sum_{k=1}^K \omega_{ik}^{(m)} \log \pi_k$$

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

تعریف می‌شود، به طوری که $\omega_{ik}^{(m)}$ امید ν_{ik} به شرط مشاهدات است و به عنوان وزن در نظر گرفته می‌شود و به صورت

$$\omega_{ik}^{(m)} = \frac{\pi_k^{(m)} \exp(y_i \log \tilde{\lambda}_k^{(m)} - \tilde{\lambda}_k^{(m)} - \log y_i!)}{\sum_{k=1}^K \pi_k^{(m)} \exp(y_i \log \tilde{\lambda}_k^{(m)} - \tilde{\lambda}_k^{(m)} - \log y_i!)}$$

قابل محاسبه است.

گام M: چون در $(m+1)$ امین مرحله از تکرار، $Q(\Psi; \Psi_{(m)})$ نسبت به عناصر بردار Ψ ماکسیم می‌شود و ماکسیم کردن $Q(\Psi; \Psi_{(m)})$ بر مبنای نسبت‌های آمیختگی به همراه دیگر پارامترها دارای پیچیدگی محاسباتی است، خلیلی و چن (۲۰۰۷) پیشنهاد کردند که در هر مرحله بر مبنای وزن $\omega_{ik}^{(m)}$ نسبت‌های آمیختگی به صورت

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

و متناسب با هر مرحله از تکرار به روز شوند. سپس با ثابت در نظر گرفتن π_k عبارت $Q(\Psi; \Psi_{(m)})$ نسبت به a, b و β ماکسیم شود. با توجه به (δ) و a و b ضرایب بخش ناپارامتری هستند که با به کارگیری الگوریتم EM به دنبال برآورد تقریبی آنها هستیم. بنابراین به منظور برآورد ضرایب بخش پارامتری

و ناپارامتری، در هر مرحله از تکرار الگوریتم، باید به حل معادله‌های زیر بپردازیم:

$$\sum_{i=1}^n \omega_{ik}^{(m)} \frac{\partial}{\partial \beta_{kj}} \{y_i \log \tilde{\lambda}_k - \tilde{\lambda}_k - \log y_i! + \log k_h(u_i - u)\} = 0 \quad (7)$$

$$\sum_{i=1}^n \omega_{ik}^{(m)} \frac{\partial}{\partial a_{kj}} \{y_i \log \tilde{\lambda}_k - \tilde{\lambda}_k - \log y_i! + \log k_h(u_i - u)\} = 0 \quad (8)$$

$$\sum_{i=1}^n \omega_{ik}^{(m)} \frac{\partial}{\partial b_{kj}} \{y_i \log \tilde{\lambda}_k - \tilde{\lambda}_k - \log y_i! + \log k_h(u_i - u)\} = 0 \quad (9)$$

همان‌طور که ملاحظه می‌شود، یک دستگاه معادلات پیچیده از مجهول‌ها ایجاد شده است که حل آن به طور تحلیلی امکان‌پذیر نیست، در نتیجه با کمک گرفتن از محیط نرم‌افزار MATLAB و با ۲۰۰۰ مرتبه تکرار، بهترین جواب همگرا به نتیجه واقعی ایجاد می‌گردد. در این مرحله به جای ضریب مجهول بخش ناپارامتری از برآورد بهینه \tilde{a} استفاده می‌شود و داریم $\tilde{a}(u) = \tilde{a}$.

مرحله دوم: محاسبه برآورد ضرایب β تاوانیده

برای برآورد ضرایب β تاوانیده، تابع درست‌نمایی تاوانیده به صورت

$$\ell(\beta) = \sum_{i=1}^n \log \sum_{k=1}^K \{ \pi_k \exp(y_i \log \lambda_k^* - \lambda_k^* - \log y_i!) \} - p_n(\Psi) \quad (10)$$

تعریف می‌شود. تفاوت λ_k^* با $\tilde{\lambda}_k$ در این است که به جای تابع مجهول ضرایب ناپارامتری مرحله قبل، از برآورد بهینه حاصل از مرحله اول استفاده می‌شود تا برآوردهای دقیق‌تری برای ضرایب پارامتری تاوانیده محاسبه شود. با ماکسیم کردن تابع درست‌نمایی تاوانیده $\ell(\beta)$ نسبت به β ، برآورد تاوانیده ضرایب بخش پارامتری به دست می‌آیند. در (۱۰) به جای $p_{nk}(\beta)$ در

$$p_n(\Psi) = \sum_{k=1}^K \pi_k \left\{ \sum_{j=1}^P p_{nk}(\beta_{kj}) \right\} \quad (11)$$

از تقریب درجه دوم موضعی آن در نزدیکی نقطه β استفاده می‌شود. دلیل استفاده از این تقریب، رفع مشکل عدم مشتق‌پذیری تابع تاوان $p_{nk}(\beta)$ در $\beta = 0$ است (فن و لی ۲۰۰۱).

گام E : این گام نیز مشابه گام محاسبه امیدریاضی مرحله اول است، با این تفاوت که امید شرطی لگاریتم تابع درست‌نمایی کامل تاوانیده $\tilde{\ell}_c(\Psi)$ به صورت $\tilde{\ell}^c(\Psi) = \ell^c(\Psi) - \tilde{p}_n(\Psi)$ به شرط متغیرهای نشانگر ν_{ik} ها و مشاهدات (u_i, x_i, z_i, y_i) محاسبه می‌شود و به صورت

$$Q(\Psi; \Psi^{(m)}) = \sum_{i=1}^n \sum_{k=1}^K \omega_{ik}^{(m)} \{y_i \log \lambda_k^* - \lambda_k^* - \log y_i!\} + \sum_{i=1}^n \sum_{k=1}^K \omega_{ik}^{(m)} \log \pi_k - \tilde{p}_n(\Psi) \quad (12)$$

است. اکنون به جای $\tilde{p}_n(\Psi)$ بر مبنای تابع‌های تاوان LASSO، SCAD و HARD، از تقریب درجه دوم موضعی استفاده می‌شود. براساس (۱۱) پس از جایگذاری تابع‌های تاوان به جای $p_{nk}(\beta_{jk})$ ، تابع‌های حاصل به دلیل وجود ساختار آمیزه‌ای به ترتیب آمیزه‌ای از انحراف مطلق به طور هموار بریده شده (SCAD)، آمیزه‌ای از کمترین انقباض مطلق و عملگر گزینش (LASSO) و آمیزه‌ای سخت^{۱۴} (HARD) خوانده می‌شوند و با نمادهای MIXSCAD، MIXLASSO و MIXHARD (خلیلی و چن ۲۰۰۷) نشان داده می‌شوند و پیشوند MIX مخفف کلمه MIXTURE است.

۱. $\tilde{p}_n(\Psi)$ به ازای تابع تاوان LASSO :

$$\tilde{p}_n^L(\Psi; \Psi^{(m)}) = \sum_{k=1}^K \pi_k \sum_{j=1}^P \{\gamma_{nk} \sqrt{n} |\beta_{jk}^{(m)}| + \frac{1}{2\beta_{jk}^{(m)}} [\sqrt{n} \gamma_{nk} \operatorname{sgn}(\beta_{jk}^{(m)})] (\beta_{jk}^2 - \beta_{jk}^{(m)2})\} \quad (13)$$

۲. $\tilde{p}_n(\Psi)$ به ازای تابع تاوان HARD :

$$\begin{aligned} \tilde{p}_n^H(\Psi; \Psi^{(m)}) &= \sum_{k=1}^K \pi_k \sum_{j=1}^P \{\gamma_{nk} - (\sqrt{n} |\beta_{jk}^{(m)}| - \gamma_{nk})^2 I(\sqrt{n} |\beta_{jk}^{(m)}| < \gamma_{nk}) \\ &+ \frac{1}{\beta_{jk}^{(m)}} \sqrt{n} (\operatorname{sgn}(\beta_{jk}^{(m)})) (\sqrt{n} \|\beta_{jk}^{(m)}\| - \gamma_{nk}) \\ &\times I(\sqrt{n} \|\beta_{jk}^{(m)}\| < \gamma_{nk}) (\beta_{jk}^2 - \beta_{jk}^{(m)2})\} \end{aligned} \quad (14)$$

^{۱۴}Mixture of Hard

۳. $\tilde{p}_n(\Psi)$ به ازای تابع تاوان SCAD :

$$\begin{aligned} \tilde{p}_n^S(\Psi; \Psi^{(m)}) &= \sum_{k=1}^K \pi_k \sum_{j=1}^P \{ \gamma_{nk} \sqrt{n} I(\sqrt{n} \|\beta_{jk}^{(m)}\| \leq \gamma_{nk}) \\ &+ \frac{\sqrt{n}(a\gamma_{nk} - \sqrt{n} \|\beta_{jk}^{(m)}\|)}{a-1} I(\sqrt{n} \|\beta_{jk}^{(m)}\| > \gamma_{nk}) \\ &+ \frac{1}{2\beta_{jk}^{(m)}} \left[\frac{-n(\text{sgn}(\beta_{jk}^{(m)}))}{a-1} I(\sqrt{n} \|\beta_{jk}^{(m)}\| > \gamma_{nk}) \right] (\beta_{jk}^2 - \beta_{jk}^{(m)2}) \} \end{aligned} \quad (15)$$

که در آن $a > 0$ و $\gamma_{nk} > 0$ پارامترهای مجهول تابع‌های مورد نظر هستند. محققانی چون فن و لی (۲۰۰۱) در مطالعات خود $a = 3.7$ در نظر گرفتند و به این نتیجه رسیدند که در شبیه‌سازی‌ها نتیجه‌های قابل قبول حاصل می‌شود. γ_{nk} ‌ها پارامترهای تنظیم‌سازی^{۱۵} نامیده می‌شوند که بر اثر تجربه محقق و یا روش‌های اعتبارسنجی متقابل^{۱۶} (CV) و اعتبارسنجی متقابل تعمیم‌یافته^{۱۷} (GCV) (کراون و واهبا ۱۹۷۹) محاسبه می‌شوند. خلیلی و چن (۲۰۰۷) یک معیار GCV بر مبنای انحراف به ازای هر مولفه برای آمیزه‌ای متناهی از مدل‌های رگرسیونی ارائه کردند. با جای‌گذاری توابع تاوان (۱۳)، (۱۴) و (۱۵) در امید شرطی و ماکسیم کردن آن در گام M نسبت به پارامترهای مجهول، برآورد ضرایب پارامتری تاوانیده مختص هریک از این نوع توابع به دست می‌آیند.

مرحله سوم: محاسبه برآورد دقیق ضرایب ناپارامتری

در این مرحله از برآوردهای تاوانیده ضرایب پارامتری $\hat{\beta}$ حاصل از مرحله قبل استفاده می‌شود و به جای برآوردهای موضعی غیرتاوانیده مرحله اول قرار داده می‌شود. در واقع در این مرحله

$$\tilde{\lambda}_k^* = \exp(x_i^T a_k + x_i^T b_k (u_i - u) + z_i^T \hat{\beta}_k)$$

به جای $\tilde{\lambda}_k$ در مرحله اول جایگزین می‌شود که در آن ضرایب بخش ناپارامتری a و b مجهول در نظر گرفته می‌شوند و برآورد پارامتری ضرایب رگرسیونی مرحله دوم جایگذاری می‌گردد. تابع درست‌نمایی در این مرحله

^{۱۵}Tuning parameter

^{۱۶}Cross-Validation

^{۱۷}Generalized Cross-Validation

بر اساس $\tilde{\lambda}_k^*$ به صورت

$$\sum_{i=1}^n \log \sum_{k=1}^K \{ \pi_k \exp(y_i \log \tilde{\lambda}_k^* - \tilde{\lambda}_k^* - \log y_i!) k_h(u_i - u) \}$$

است. در نتیجه تابع درستنمایی کامل مانند (۵) بر مبنای λ_k^* تعریف می‌شود.

گام E: امید $\ell_n^c(\Psi)$ به شرط متغیرهای نشانگر نامشهود ν_{ik} و مشاهدات (u_i, x_i, z_i, y_i) به صورت

$$\begin{aligned} Q(\Psi; \Psi^{(m)}) &= \sum_{i=1}^n \sum_{k=1}^K \omega_{ik}^{(m)} \{ y_i \log \tilde{\lambda}_k^* - \tilde{\lambda}_k^* - \log y_i! \} + \sum_{i=1}^n \sum_{k=1}^K \omega_{ik}^{(m)} \log \pi_k \\ &+ \sum_{i=1}^n \sum_{k=1}^K \omega_{ik}^{(m)} \log k_h(u_i - u) \end{aligned}$$

گام M: در این گام بر مبنای $(m+1)$ امین مرحله تکرار، $Q(\Psi; \Psi^{(m)})$ نسبت به ضرایب مجهول بخش ناپارامتری، a و b ماکسیم می‌شود که بعد از به‌روزرسانی نسبت‌های آمیزه‌ای انجام می‌پذیرد. بنابراین با ثابت در نظر گرفتن π_k و با حل معادلات مانند (۷) و (۸) برآوردهای مورد نظر به دست می‌آیند. در واقع بعد از انجام محاسبات گام M در مرحله دوم، به منظور برآورد دقیق‌تر ضرایب مجهول بخش ناپارامتری در مقابل برآوردهای موضعی ضرایب ناپارامتری در مرحله اول، به حل معادلات با در نظر گرفتن $\tilde{\lambda}_k^*$ به جای $\tilde{\lambda}_k$ پرداخته می‌شود. در اینجا نیز همانند مرحله اول با کمک گرفتن از محیط نرم‌افزار معرفی شده، جواب بهینه دستگاه معادلات با تعداد تکرار معلوم به دست آورده می‌شود. بنابراین در مرحله سوم برآوردهای دقیق ضرایب ناپارامتری و از مرحله دوم برآوردهای تاوانیده ضرایب پارامتری به صورت $\{\hat{a}, \hat{b}, \hat{\beta}\}$ حاصل می‌شوند. روند انجام محاسبات به طور کلی در پنج گام زیر انجام می‌پذیرد:

۱. در نظر گرفتن یک مقدار اولیه برای β تحت عنوان β_0 .

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

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

۴. 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} .

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

۴ مطالعه شبیه‌سازی

به منظور انجام شبیه‌سازی برای مراحل مطرح شده، مقادیر اولیه به صورت زیر در نظر گرفته می‌شوند: $k = 3$ در نظر گرفته می‌شود، به این ترتیب سه جامعه با نسبت‌های آمیختگی $\pi_1 = 0.5$ ، $\pi_2 = 0.3$ و $\pi_3 = 0.2$ فرض می‌شود. فرض کنید X متغیری با توزیع نرمال دو متغیره به صورت $X \sim N_2(\mu, \Sigma)$ باشد و به همین ترتیب Z نیز متغیری با توزیع نرمال ده متغیره به صورت $Z \sim N_{10}(\mu, \Sigma)$ فرض می‌شود، با میانگین صفر و ماتریس کوواریانس که اجزای آن از طریق رابطه σ^{i-j} همانند Σ متغیر X محاسبه می‌شود. همچنین مقادیر اولیه ضرایب β بخش پارامتری در جدول ۲ نشان داده شده است.

همانند لی و لیانگ (۲۰۰۸) و اسکندری و اورمز (۲۰۱۵) در ارتباط با بخش ناپارامتری و تابع ناپارامتری $\alpha(\cdot)$ با استفاده از یک تابع غیر خطی دلخواه مانند توابع نمایی یا مثلثاتی، تقریبی برای تابع ناپارامتری به صورت $\alpha_{k1} = \frac{1}{4} \exp(3u + 1)$ و $\alpha_{k2} = 4u(1 + u^2)$ در نظر گرفته می‌شود که در آن U متغیری با توزیع $U(0, 1)$ است.

بر مبنای اطلاعات اولیه مطرح شده، ۱۰۰ داده از سه جامعه با توزیع‌های $Poisson(\lambda_1)$ ، $Poisson(\lambda_2)$ و $Poisson(\lambda_3)$ تولید می‌کنیم که به عنوان نمونه λ_1 به شکل زیر محاسبه می‌شود:

$$\lambda_1 = \exp\left(x^T \begin{pmatrix} \frac{1}{4} \exp(3u + 1) \\ 4u(1 + u^2) \end{pmatrix} + z^T \beta\right)$$

مقدار میانگین و واریانس بر مبنای داده‌های شبیه‌سازی شده از توزیع پواسون به ترتیب عبارتند از $10/532$ و $896/705$. به علت وجود صفرهای زیاد در داده‌های تولید شده بر مبنای توزیع پواسون، مشاهده می‌کنید که مقدار واریانس بسیار بزرگ است. بنابراین پراکندگی بیش از حد رخ داده است و همانطور که قبلاً نیز توضیح داده شد، به منظور محاسبه برآوردهایی با دقت بیشتر به جای توزیع پواسون از توزیع دوجمله‌ای منفی برای تولید داده‌ها استفاده می‌کنیم. در نتیجه بر مبنای تعریف تابع توزیع دوجمله‌ای منفی در مدل‌های آمیزه‌ای (زوو و همکاران، ۲۰۱۳)

$$Y \sim NB(r, p), \quad p_k = \frac{\mu_k}{\mu_k + \phi_k}, \quad \mu_k = \exp(x^T \alpha_k(u) + z^T \beta_k) \quad k = 1, 2, 3$$

را تولید می‌کنیم. بر مبنای توابع ناپارامتری، ضرایب پارامتری و بقیه موارد همانند ابتدای بخش ۴ در نظر گرفته شده‌اند.

۴,۱ بررسی عملکرد روش ارائه‌شده در ارتباط با بخش پارامتری

به منظور ارزیابی شیوه انتخاب متغیر مربوط به بخش پارامتری از شاخص میانگین توان دوم خطای تعمیم‌یافته^{۱۸} (GMSE) استفاده می‌شود. در واقع GMSE بر اساس برآورد ضرایب تاوانیده نهایی که از ماکسیم‌سازی (۱۲) و پس از طی گام‌های E و M به دست می‌آید، به صورت زیر محاسبه می‌شود:

$$GMSE(\hat{\beta}) = E[Z^T(\hat{\beta} - \beta)]^2 = (\hat{\beta} - \beta)E(ZZ^T)(\hat{\beta} - \beta)$$

در جدول ۱ با مقایسه مقادیر GMSE بر مبنای دو توزیع متفاوت، مشاهده می‌شود که با در نظر گرفتن

جدول ۱. مقدار شاخص GMSE بر مبنای تابع‌های تاوان متفاوت

تابع تاوان			
MIXSCAD	MIXHARD	MIXLASSO	
۰/۳۴۰	۰/۵۶۱	۰/۷۰۳	پواسون
۰/۱۸۱	۰/۲۱۲	۰/۵۷۴	دوجمله‌ای منفی

توزیع دوجمله‌ای منفی، این مقدار کاهش یافته است. از سوی دیگر شیوه MIXSCAD با کمترین میزان GMSE در هر دو توزیع از عملکرد بهتری برخوردار است، چون میزان خطا در برآورد ضرایب رگرسیونی تاوانیده با استفاده از تابع تاوان MIXSCAD نسبت به هر یک از دو تابع تاوان MIXHARD و MIXLASSO کمتر است. در جدول ۲ میانگین مقدار برآورد ضرایب به همراه انحراف استانداردشان به ازای ۲۰۰۰ مرتبه اجرای برنامه نشان داده می‌شود. در واقع هدف برآورد ضرایب صفر به درستی بوده است و همانطور که مشاهده می‌شود این امر تا حد قابل قبولی به درستی تحقق یافته است چون میانگین برآورد ضرایب صفر بسیار کوچک و نزدیک صفر شده‌اند. اما تعدادی از ضرایب صفر به اشتباه مخالف صفر برآورد شده‌اند، چون میانگین برآوردها عددی مخالف صفر ولی نزدیک به آن حاصل شده است. با این وجود اثر این برآوردهای نادرست در طول اجرای تمام مرحله‌ها قابل چشم‌پوشی است، به دلیل اینکه میانگین کل مقادیر به طور تقریبی برابر با صفر است.

همان‌طور که در جدول ۲ ملاحظه می‌شود برآورد تاوانیده حاصل از تابع تاوان MIXSCAD از دقت بیشتری برخوردار است. برآوردهای حاصل از تاوان MIXHARD همانند برآوردهای MIXSCAD

^{۱۸}Generalized Mean Square Error

جدول ۰۲. برآورد ضرایب پارامتری و انحراف معیار آنها بر مبنای ۲۰۰۰ مرتبه تکرار

MIXLASSO		MIXHARD		MIXSCAD		β	K
SD	$\hat{\beta}$	SD	$\hat{\beta}$	SD	$\hat{\beta}$		
۰/۱۶۷۱	۲/۸۸۱۲	۰/۱۶۸۲	۲/۷۵۲۱	۰/۱۵۲۳	۲/۶۶۲۱	۳	
۰/۱۳۲۲	۰/۹۸۰۱	۰/۱۳۳۱	۰/۸۸۲۳	۰/۱۴۳۱	۰/۷۰۲۲	۱	
۰/۱۴۱۳	-۰/۰۳۰۱	۰/۱۴۳۴	-۰/۰۴۵۱	۰/۱۶۳۴	-۰/۰۶۶۱	۰	
۰/۱۴۸۱	۱/۷۸۳۲	۰/۱۴۸۲	۱/۶۱۴۲	۰/۱۶۶۲	۱/۴۶۲۲	۲	
۰/۱۶۱۲	۰/۰۲۱۰	۰/۱۶۳۳	۰/۰۳۳۵	۰/۱۷۰۲	۰/۰۵۸۱	۰	۱
۰/۱۳۴۲	۰/۰۴۳۲	۰/۱۳۵۱	۰/۰۴۶۳	۰/۱۵۳۱	۰/۰۶۱۲	۰	
۰/۱۱۹۱	۰/۰۸۶۲	۰/۱۱۸۲	۰/۰۷۰۹	۰/۱۲۴۰	۰/۰۷۱۹	۱	
۰/۱۸۵۳	۱/۴۳۶۰	۰/۱۸۸۵	۱/۳۰۴۱	۰/۱۹۰۳	۱/۲۹۸۰	۱/۵	
۰/۱۷۳۳	-۰/۰۴۵۱	۰/۱۶۷۷	-۰/۰۵۴۴	۰/۱۸۲۶	-۰/۰۵۶۲	۰	
۰/۱۶۳۱	۱/۷۵۲۰	۰/۱۶۵۱	۱/۷۳۲۲	۱/۸۶۳۳	۱/۵۳۲۱	۲	
۰/۳۲۱۲	۲/۸۸۷۴	۰/۳۲۳۱	۲/۷۴۴۱	۰/۳۸۸۲	۲/۶۰۱۲	۲	
۰/۳۰۳۱	۱/۰۹۶۲	۰/۳۰۵۱	۱/۱۱۲۳	۰/۳۸۸۷	۱/۳۰۳۲	۱	
۰/۳۱۱۲	۰/۰۱۲۱	۰/۳۰۴۲	۰/۰۱۵۱	۰/۳۵۵۶	۰/۰۲۸۳	۰	
۰/۳۲۲۸	۰/۰۱۰۸	۰/۳۱۱۴	۰/۰۲۱۳	۰/۳۴۳۳	۰/۰۲۲۴	۰	
۰/۳۰۴۴	۱/۳۸۹۸	۰/۳۰۲۱	۱/۳۸۱۲	۰/۳۳۶۲	۱/۲۲۶۱	۱/۵	۲
۰/۳۲۵۱	-۰/۰۰۲۷	۰/۳۲۶۷	-۰/۰۰۳۱	۰/۳۷۲۶	-۰/۰۳۱۱	۰	
۰/۳۲۸۱	۰/۰۸۶۶	۰/۳۲۵۲	۰/۰۸۷۱	۰/۳۹۵۷	۰/۰۵۲۲	۱	
۰/۳۱۶۲	۲/۷۶۶۲	۰/۳۱۱۲	۲/۶۵۳۳	۰/۳۳۴۸	۲/۶۰۴۳	۳	
۰/۳۰۹۱	۰/۰۲۱۱	۰/۳۰۷۲	۰/۰۲۲۷	۰/۳۲۹۴	۰/۰۴۱۲	۰	
۰/۳۳۲۵	۰/۰۱۹۸	۰/۳۲۲۵	۰/۰۲۲۵	۰/۳۶۵۲	۰/۰۴۸۴	۰	
۰/۱۸۳۵	۱/۴۸۲۲	۰/۱۸۸۶	۱/۴۸۹۵	۰/۱۸۹۰	۱/۴۶۰۲	۱/۵	
۰/۱۱۹۵	۳/۱۰۹۳	۰/۱۱۹۱	۳/۱۱۰۴	۰/۱۳۰۴	۳/۲۳۳۹	۳	
۰/۱۳۳۱	۱/۱۱۵۲	۰/۱۰۵۲	۱/۱۱۰۷	۰/۱۶۷۲	۱/۲۱۱۲	۱	۳
۰/۱۲۲۵	-۰/۰۱۳۴	۰/۱۲۰۲	-۰/۰۱۲۸	۰/۱۴۵۴	-۰/۰۲۲۳	۰	
۰/۱۴۵۶	-۰/۰۲۱۲	۰/۱۴۴۳	-۰/۰۲۱۳	۰/۱۵۷۲	-۰/۰۴۰۱	۰	
۰/۱۲۸۵	۱/۸۷۷۰	۰/۱۲۹۱	۱/۸۹۲۱	۰/۱۴۶۳	۱/۷۸۱۱	۲	
۰/۱۴۰۵	-۰/۰۳۱۱	۰/۱۴۱۵	-۰/۰۲۹۹	۰/۱۶۲۶	-۰/۰۳۹۸	۰	
۰/۱۳۲۸	-۰/۰۱۸۱	۰/۱۳۳۲	-۰/۰۱۸۶	۰/۱۵۱۷	-۰/۰۲۵۲	۰	
۰/۱۱۵۸	۰/۸۹۲۲	۰/۱۱۶۵	۰/۸۶۰۱	۰/۱۱۸۱	۰/۷۵۴۴	۱	
۰/۱۵۰۳	۲/۷۹۰۴	۰/۱۵۴۲	۲/۷۹۹۲	۰/۱۶۸۶	۲/۶۵۱۲	۳	

هستند اما با توجه به مقدار انحراف استانداردشان از دقت کمتری برخوردارند. بنابراین برآوردهای حاصل از شیوه MIXLASSO قابل اطمینان نیستند چون نسبت به دو تابع تاوان دیگر از دقت کمتری برخوردار می‌باشند.

۴,۲ بررسی عملکرد روش ارائه‌شده در ارتباط با بخش ناپارامتری

به منظور ارزیابی نحوه عملکرد شیوه ارائه‌شده برای برآورد بخش ناپارامتری $\hat{\alpha}(u_l)$ ، جذر متوسط توان دوم خطاها^{۱۹} (RASE) به صورت

$$RASE = \left\{ n_{grid}^{-1} \sum_{\ell=1}^{n_{grid}} \|\hat{\alpha}(u_\ell) - \alpha(u_\ell)\|^2 \right\}^{\frac{1}{2}} \quad (۱۶)$$

تعریف می‌شود، که در آن $\{u_\ell, \ell = 1, \dots, n_{grid}\}$ نقاط شبکه‌ای هستند و تابع‌های $\{\hat{\alpha}(\cdot)\}$ در این نقاط برآورد می‌شوند. در شبیه‌سازی از تابع هسته اپانشینکف^{۲۰} به صورت $K_{0.125}(u) = 6 \left(1 - \left(\frac{u_i - u}{0.125} \right)_+^2 \right)$ و بر مبنای $n_{grid} = 200$ استفاده می‌شود، به طوری که پهنای باند $h = 0.125$ در نظر گرفته شده است. با استفاده از شاخص RASE، می‌توان میزان دقت برآورد ضرایب بسط تیلور a و b را در ارتباط با بخش ناپارامتری بررسی نمود، به همین منظور $RASE_1$ و $RASE$ محاسبه می‌شوند.

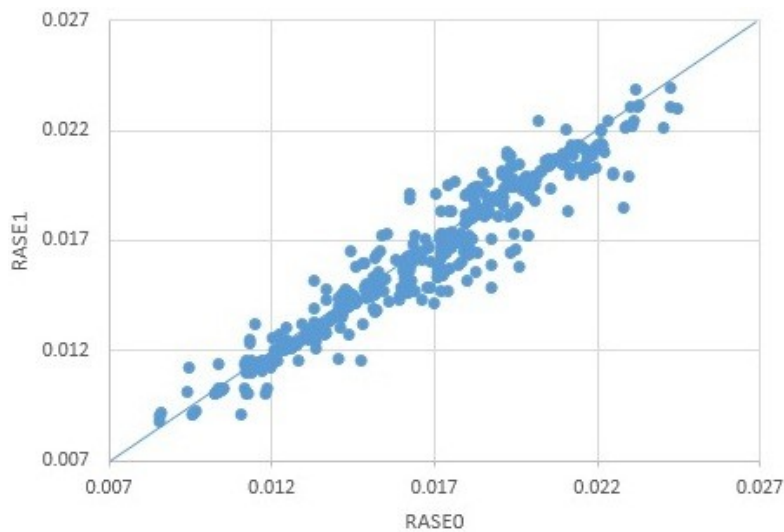
$RASE$: برآورد $\hat{\alpha}(u_l)$ با توجه به مقدار واقعی β های اولیه انجام می‌پذیرد. در واقع بر مبنای رابطه $\mu_k = \exp(x^T \alpha_k(u) + z^T \beta_k)$ و مقدارهای اولیه تعریف شده β ، بردار x ، بردار z و بر مبنای مقدار مشخصی از نرخ میانگین مولفه k ام (μ_k) مقدار تابع $\hat{\alpha}(u_\ell)$ محاسبه می‌شود. از سوی دیگر $\alpha(u_\ell)$ با توجه به $(\frac{1}{2} \exp(3u + 1), 4u(1 + u^2))$ بر مبنای مقادیر تصادفی u از بازه $(0, 1)$ انتخاب می‌شود و به این ترتیب با توجه به (۱۶)، $RASE$ محاسبه می‌شود.

$RASE_1$: روند محاسبات همانند $RASE$ است با این تفاوت که به جای $\hat{\alpha}(u_l)$ برآورد آن بر مبنای ضرایب بسط تیلور a و b ، حاصل از الگوریتم EM جایگذاری می‌شود. پس مقادیر $RASE$ و $RASE_1$ را در مقابل یکدیگر رسم می‌کنیم. این کار به منظور ارزیابی میزان دقت برآورد ضرایب بسط تیلور در تابع ناپارامتری نسبت به حالتی است که تابع ناپارامتری از ابتدا معلوم فرض شده باشد.

همان‌طور که در شکل ۱ ملاحظه می‌شود، مقادیر برآورد بر اساس ۲۰۰۰ مرتبه تکرار، به‌طور تقریبی بر

^{۱۹}Root of Average Square Errors

^{۲۰}Epanechnikov



شکل ۱. مقدار $(RASE_0)$ در مقابل $(RASE_1)$

روی خط قرار گرفته‌اند، به این معنی که در برآورد ضرایب بخش ناپارامتری، عملکرد الگوریتم EM تقریباً به درستی زمانی بوده است که محاسبات به طور مستقیم با استفاده از مقادیر حقیقی معلوم β انجام می‌پذیرد، یا به عبارت دیگر حاکی از برابری تقریبی $(RASE_0)$ و $(RASE_1)$ است.

بحث و نتیجه‌گیری

تاکنون موضوع انتخاب متغیر در آمیزه‌ای متناهی از مدل‌های رگرسیونی (خلیلی و چن ۲۰۰۷) و در مدل‌های رگرسیونی نیم‌پارامتری (لی و لیانگ، ۲۰۰۸) و همچنین مطالعه ترکیبی این دو حالت با تکیه بر روی بخش ناپارامتری مورد بررسی قرار گرفته است. به طوری که در مدل اسکندری و اورمز (۲۰۱۵) حالت تعمیم‌یافته مدل پیوند بررسی شده است. اما مدلی مد نظر قرار گرفت که در آن داده‌ها از توزیع پواسون تبعیت می‌کنند و گاهی شرایطی پیش می‌آید که به علت وجود صفرهای بیش از حد و بیش‌پراکندگی، نتیجه برآورد پارامترها خارج از انتظار است و قابل قبول نخواهند بود. در چنین شرایطی استفاده از توزیع دوجمله‌ای منفی به جای توزیع پواسون توصیه می‌شود. هدف ما بررسی درستی برآوردهای حاصل از مدلی است که داده‌های آن برای تصحیح مشکل درستی برآوردها از توزیع دوجمله‌ای منفی تبعیت می‌کنند. در بررسی درستی برآوردها در بخش پارامتری بر اساس ملاک GMSE ملاحظه شد که مقدار خطا در

برآورد ضرایب پارامتری به دلیل استفاده از توزیع دوجمله‌ای منفی کاهش می‌یابد. علاوه بر این برآوردهای حاصل از تابع تاوان MIXSCAD از دقت بیشتری نسبت به دو تابع تاوان دیگر برخوردار بودند. از سوی دیگر مقدار میانگین برآورد نقاط صفر پس از چندین مرتبه اجرای الگوریتم، مقداری کوچک و نزدیک به صفر به دست آمد، بنابراین نقاط صفر به درستی برآورد شده‌اند. همچنین برآوردهای حاصل از تابع تاوان MIXSCAD نسبت به دو تابع تاوان دیگر از دقت بیشتر و انحراف استاندارد کمتری برخوردار است. لازم به ذکر است که تابع تاوان MIXHARD تقریباً رفتاری مشابه با MIXSCAD دارد. در بررسی عملکرد برآورد ضرایب ناپارامتری به هنگام استفاده از توزیع دوجمله‌ای منفی براساس ملاک RASE ملاحظه شد که میزان دقت برآورد ضرایب ناپارامتری در حضور ضرایب تاوانیده به خوبی زمانی بوده است که توابع ناپارامتری از ابتدا معلوم فرض شده بودند. بنابراین می‌توان نتیجه گرفت در مواقعی که حالت بیش‌پراکندگی یا وجود صفرهای بیش از حد به هنگام استفاده از توزیع پواسون رخ می‌دهد، توزیع دوجمله‌ای منفی می‌تواند به منظور برآورد صحیح پارامترها و کسب نتیجه‌ای مطلوب کارآمد واقع شود.

تقدیر و تشکر

نویسندگان از مسئولان مجله علوم آماری در فراهم کردن زمینه همکاری علمی برای موضوعات جدید آماری تشکر می‌نمایند و از داوران محترم برای ارائه پیشنهادات کارآمد در بهبود کیفی این مقاله کمال سپاس و قدردانی را دارند.

مراجع

- Craven, P. and Wahba, G. (1979), Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation, *Numerische Mathematika*, **31**, 377 – 403.
- Eskandari, F. and Ormoz, E. (2015), Finite Mixture of Generalized Semi-parametric Models: Variable Selection via Penalized Estimation, *Communications in Statistics-Simulation and Computation*, **45**, 3744-3759.
- Fan, J. and Li, R. (2001), Variable Selection via Penalized Likelihood and It's Oracle Properties, *Journal of American Statistical Association*, **6**, 1348-1360.
- Hardle, W., Liang, H. and Gao, J. (2000), *Partially Linear Models: Contributions to Statistics*, Physica-Verlag, Heidelberg.

- Hunter, D. R. and Young, D. S. (2012), Semi-Parametric Mixtures of Regressions, *Journal of Non-parametric Statistics*, **24**, 19-38.
- Khalili, A. and Chen, J. (2007), Variable Selection in Finite Mixture of Regression Models, *Journal of American Statistical Association*, **102**, 1025-1038.
- Lambert, D. (1992), Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing, *Technometrics*, **34**, 1-14.
- Li, R. and Liang, H. (2008), Variable Selection in Semi-Parametric Regression Modeling, *Annual Statistics*, **36**, 261-286.
- Lim, H. K., Li, W. K. and Yu, P. L. H. (2014), Zero-Inflated Poisson Regression Mixture Model, *Computational Statistics and Data Analysis*, **71**, 151-158.
- McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, John Wiley, New York.
- Ormoz, E. and Eskandari, F. (2016), Variable Selection in Finite Mixture of Semi-Parametric Regression Models, *Communications in Statistics-Theory and Methods*, **45**, 695-711.
- Ruppert, R., Wand, M. and Carroll, R. (2003), *Semi-Parametric Regression*, Cambridge University Press, Cambridge.
- Tibshirani, R. (1996), Regression Shrinkage and Selection via the LASSO, *Journal of Royal Statistical Society*, **58**, 267-288.
- Yatchew, A. (2003), *Semi-Parametric Regression for the Applied Econometrician*, Cambridge University Press, Cambridge.
- Zou, Y., Zhang, Y. and Lord, D. (2013), Application of Finite Mixture of Negative Binomial Regression Models with Varying Weight Parameters for Vehicle Crash Data Analysis, *Accident Analysis and Prevention*, **50**, 1042-1051.