مجله مدل سازی پیشرفته ریاضی
سال ۱۳۹۹، دوره ۱۰، شماره ۲، ص ۳۹۹-۳۷۹
شاپا چاپی: ۸۰۸۸-۲۲۵۱
DOI: 10.22055/JAMM.2020.31201.1763

مدل رگرسیون نیمه‌پارامتری مکان-مقیاس با دم‌های نیمه‌سنگین بر اساس توزیع هایپربولیک سکانت

جمیل اونق، حسین باغیشنی[1]، احمد نزاکتی

گروه آمار، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود

تاریخ دریافت: ۱۳۹۸/۷/۴ تاریخ پذیرش: ۱۳۹۹/۳/۳

چکیده: کاربران مدل‌های رگرسیون کلاسیک دریافته‌اند که در عمل بسیاری از پذیره‌های این نوع مدل‌ها برقرار نیستند و باید مدل‌هایی را به کار گرفت که قادر به مدل‌بندی ماهیت واقعی داده‌ها باشند. رده مدل‌های جمعی تعمیم‌یافته برای همه پارامترهای یک توزیع شامل مکان، مقیاس و شکل، یک رده بسیار منعطف و پرطرفدار است که می‌تواند پیچیدگی‌های موجود در داده‌ها را لحاظ کند. در کنار ارائه یک مدل رگرسیونی برای پارامترهای مختلف توزیع متغیر پاسخ و نه‌فقط میانگین، مدل‌بندی داده‌های پرت نیز دارای اهمیت است. در مواردی که تعداد داده‌های پرت اندک است، استفاده از توزیع‌های دم‌سنگین می‌تواند پیچیدگی بیش‌ازحد نیاز وارد مسئله کند. در این مقاله، با در نظر گرفتن توزیع هایپربولیک سکانت با دم نیمه‌سنگین و تعبیه آن در چارچوب مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل، یک مدل رگرسیون نیمه‌پارامتری مکان-مقیاس جدید را برای رفع این مشکل در کنار حفظ انعطاف بالای مدل‌بندی اثرات متغیرهای رگرسیونی، معرفی می‌کنیم. کارایی مدل پیشنهادی را در مقایسه با مدل کلاسیک نرمال با یک مطالعه شبیه‌سازی بررسی می‌کنیم و کاربست آن را در یک مثال واقعی نمایش می‌دهیم.

واژه‌های کلیدی: توزیع با دم نیمه‌سنگین، توزیع هایپربولیک سکانت، داده پرت، درستنمایی تاوانیده، رگرسیون مکان-مقیاس.

رده‌بندی ریاضی (۲۰۱۰): ۶۲G۰۸، ۶۲J۰۵.

۱-آدرس الکترونیکی نویسنده مسوول، آدرس: hbaghishani@shahroodut.ac.ir

۱- مقدمه

رشد فنّاوری در کنار مزایای متعدد، منجر به تولید مجموعه داده‌های با پیچیدگی‌های بیشتر شده است. تعمیم‌های رو به گسترش مدل‌های آماری و روش‌های برازش آن‌ها، در هر دو دیدگاه کلاسیک و بیزی، بستری برای مدل‌بندی واقعی‌تر این نوع داده‌ها فراهم آورده‌اند.

مدل‌های رگرسیونی یکی از مهم‌ترین زیررده‌های مدل‌های آماری هستند که در اغلب زمینه‌های علمی، مانند اقتصاد، پزشکی، علوم اجتماعی، هواشناسی و صنعت، مورداستفاده قرار می‌گیرند. در چارچوب مدل‌بندی رگرسیونی، مدل‌های خطی تعمیم‌یافته[1] (GLMs، نلدر و ودربرن [۱]) و مدل‌های جمعی تعمیم‌یافته[2] (GAMs، هیستی و تیبشیرانی [۲]) جایگاه ویژه‌ای دارند. در هر دوی این زیررده‌ها توزیع متغیر پاسخ، y ، عضوی از خانواده توزیع‌های نمایی در نظر گرفته می‌شود، ه در آن میانگین پاسخ، μ ، به‌عنوان تابعی از متغیرهای تبیینی رگرسیونی مدل‌بندی می‌شود و واریانس پاسخ به پارامتر ثابت پراکنش[3] φ و میانگین μ از طریق تابع واریانس $\tau(\mu)$ به‌صورت $V(y)=\varphi\tau(\mu)$ وابسته است؛ بنابراین، ر مدل‌های GLM و GAM واریانس توزیع پاسخ به‌صورتی واضح و سرراست برحسب متغیرهای تبیینی مدل‌بندی نمی‌شود و تنها از طریق میانگین به‌صورت تلویحی به آن‌ها وابسته است. توجه داشته باشید که مدل‌های رگرسیون خطی نیز عضوی از GLMs هستند، به‌طوری‌که برای آن‌ها $\tau(\mu)=1$.

اگرچه مدل‌های GAM با در نظر گرفتن اثرات غیرخطی (ناپارامتری) متغیرهای تبیینی به‌صورت جمعی، نسبت به مدل‌های GLM، انعطاف بیشتری دارند، اما به‌غیراز پارامتر میانگین (مکان) اجازه مدل‌بندی سایر پارامترها، مثل واریانس توزیع پاسخ را نمی‌دهند. برای رفع این محدودیت، ریگبی و استاسینوپولوس [۳] رده مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل[4] (GAMLSS) را معرفی کردند. دو ویژگی مهم این رده از مدل‌ها عبارتند از:

۱. یک خانواده کلی از توزیع‌های پارامتری جایگزین خانواده توزیع‌های نمایی برای متغیر پاسخ می‌شود.

۲. نه‌فقط میانگین (مکان) بلکه همه پارامترهای توزیع شرطی پاسخ را می‌توان به‌عنوان تابعی پارامتری یا ناپارامتری جمعی از متغیرهای تبیینی مدل‌بندی کرد.

1- Generalized Linear Models (GLMs)
2- Generalized Additive Models (GAMs)
3- Dispersion
4- Generalized Additive Models for Location, Scale, and Shape (GAMLSS)

لازم به ذکر است که برای تحلیل پاسخ‌های وابسته، معمولاً از نسخه‌های آمیخته[1] مدل‌های GLM (GLMM)، بریسلو و کلیتون [۴]؛ لی و نلدر [۵]) و GAM (GAMM)، لین و زانگ [۶]؛ فاهرمیر و لانگ [۷]) استفاده می‌شود که در آن‌ها با افزودن اثرات تصادفی به‌صورت جمعی به پیشگوی مدل، ساختار وابستگی داده‌ها وارد مدل می‌شود. رده مدل‌های GAMLSS نیز این ویژگی را دارا است که اثرات تصادفی را به هرکدام از پارامترهای توزیع اضافه کند.

در دنیای واقعی، همیشه داده‌ها رفتاری ایدئال برای مدل‌بندی ندارند و معمولاً یک یا چند داده مشکوک (پرت) در جمع داده‌ها حضور دارند که می‌توانند بر الگوی حاکم بر داده‌ها تأثیر زیادی داشته باشند. در رده مدل‌های GAMLSS، توزیع‌های پارامتری مختلفی هستند که به‌طور ذاتی دم‌سنگین و قادر به مدل‌بندی داده‌های پرت توسط پارامترهای مقیاس و شکل می‌باشند (استاسینوپولوس و همکاران [۸]). به‌عنوان‌مثال می‌توان به توزیع‌های چوله-نرمال و چوله-t اشاره کرد. علی‌رغم انعطاف قابل‌توجه مدل‌های موجود در رده GAMLSS که در دو بسته‌افزار *gamlss* (استاسینوپولوس و ریگبی [۹]) و *gamlss.dist*[2] در محیط نرم‌افزار *R* وجود دارند، امکان رخداد دو مشکل بالقوه وجود دارد:

۱. در ساختار زیررده مدل‌های رگرسیون خطی، معمولاً متقارن بودن جمله خطا با میانگین صفر پذیرفته می‌شود. این پذیره برای توزیع‌های چوله برقرار نیست و بنابراین استفاده از برخی از اعضای رده GAMLSS برای حالت خاص مدل‌های خطی منطقی نیست.

۲. زمانی‌که تعداد داده‌های پرت نسبت به کل حجم نمونه کوچک است، توزیع‌های دم‌سنگین متقارن می‌توانند تأکید بیش‌ازحد لازم را به دم‌ها اختصاص دهند و در مقابل چگالی کمتری را برای مقادیر مرکزی توزیع (که اغلب داده‌های موجود را شامل می‌شوند) در نظر بگیرند. این پدیده می‌تواند به مدلی نامناسب منتهی شود.

با توجه به این دو مشکل، معرفی و امکان به‌کارگیری یک توزیع متقارن با دم‌های نیمه‌سنگین در رده GAMLSS می‌تواند کارساز باشد.

هدف اصلی ما در این مقاله، معرفی یک مدل GAMLSS جدید کارا در برخورد با پاسخ‌های با دم نیمه‌سنگین بر پایه توزیع هایپربولیک سکانت[3] (HS)، به‌عنوان یک توزیع متقارن با دم نیمه‌سنگین، برای ایجاد انعطاف لازم در مدل‌بندی رگرسیونی و برخورد با نقاط پرت با تعداد کم نسبت به سایر داده‌ها است. انتظار داریم این توزیع در مواقعی که دم توزیع جمله خطا بین دم‌های یک توزیع دم‌سنگین و نرمال باشد، عملکرد مناسبی داشته باشد. در واقع نیمه‌سنگین بودن دم

1- Mixed models
2- https://cran.r-project.org/package=gamlss.dist
3- Hyperbolic Secant

توزیع خطا به معنی امکان مدل‌بندی تعداد نه‌چندان زیاد از داده‌های پرت است. با توجه به این‌که خانواده توزیع‌های HS دارای دو پارامتر مکان و مقیاس است، مدل رگرسیون معرفی‌شده یک مدل مکان-مقیاس محسوب می‌شود. از اولین کارهایی که در آن مدل‌بندی هم‌زمان مکان و مقیاس پیشنهاد شده است، می‌توان به ریگبی و استاسینوپولوس [۱۰] اشاره کرد.

در ادامه، در بخش ۲ خانواده توزیع‌های HS و ویژگی‌های مدنظر را معرفی می‌کنیم. در بخش ۳ به معرفی و ساخت یک مدل GAMLSS جدید با توزیع خطای HS می‌پردازیم. در این بخش برازش مدل در چارچوب استنباط مبتنی بر درستنمایی تاوانیده[1] را نیز معرفی خواهیم کرد. در بخش ۴ با یک مطالعه شبیه‌سازی عملکرد مدل معرفی‌شده را در مقایسه با مدل مبتنی بر نرمال ارزیابی می‌کنیم. در بخش ۵ کاربست مدل را در یک مثال واقعی برای بررسی وضعیت اسیدی بودن دریاچه‌های واقع در رشته‌کوه تیغه آبی[2] در کشور آمریکا (داگلاس و دلامپادی [۱۱]) نشان می‌دهیم. در پایان نیز نتیجه‌گیری خواهیم کرد.

۲- خانواده توزیع‌های HS

توزیع HS اولین بار توسط فیشر [۱۲] معرفی شد. این توزیع حول پارامتر مکان خود که میانگین توزیع هم هست، متقارن است. برخی از ویژگی‌های این خانواده توسط دود [۱۳] و پرک [۱۴] موردبررسی قرار گرفتند. وگان [۱۵] این توزیع را به حالت چوله تعمیم داد و فیشچر و وگان [۱۶] توزیع هایپربولیک سکانت تعمیم‌یافته چوله را معرفی کردند. کتاب فیشچر [۱۷] با عنوان *توزیع‌های* HS *تعمیم‌یافته و کاربردهای آن در داده‌های مالی*، منبع جامع و کاملی برای مبانی نظری و کاربردهای این توزیع است. استفاده از این توزیع در مدل‌بندی و تحلیل داده‌های مالی، معمول است.

تابع چگالی توزیع HS به‌صورت

$$f_Y\left(y;\mu,\sigma\right)=\frac{1}{\pi\sigma}\frac{2exp(\frac{y-\mu}{\sigma})}{exp(\frac{2(y-\mu)}{\sigma})+1},\quad y\in R$$

تعریف می‌شود که در آن پارامتر μ میانگین توزیع و σ پارامتر مقیاس هستند. واریانس این توزیع $\frac{\pi^2}{4}\sigma^2$ است و برای نمایش توزیع از نماد $Y\sim HS\left(\mu,\sigma\right)$ استفاده می‌کنیم. این توزیع نسبت به توزیع نرمال دم‌های سنگین‌تری دارد ولی به سنگینی توزیع‌هایی مثل کوشی، لاپلاس

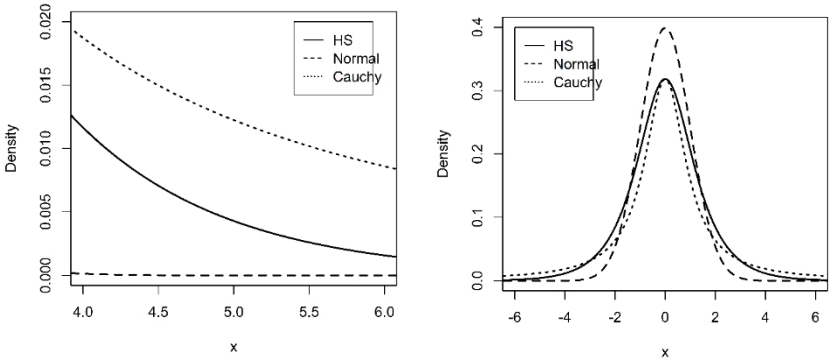
1- Penalized likelihood
2- Blue Ridge mountains

یا t-استودنت نیست؛ بنابراین؛ در متون مربوطه از آن به‌عنوان یک توزیع با دم نیمه‌سنگین یاد می‌شود. نیمه‌سنگینی دم یک توزیع البته به‌طور رسمی تعریف و برای توزیع HS اثبات می‌شود. به‌عنوان نمونه برای مشاهده تعریفی رسمی از دم‌سنگینی می‌توانید ساتو [۱۸] را ببینید.

قضیه ۱. توزیع HS یک توزیع با دم نیمه‌سنگین است.

اثبات: برای مشاهده برهان به ساتو [۱۸] مراجعه کنید.

برای درک نیمه‌سنگینی دم توزیع HS، شکل (۱) منحنی توابع چگالی و دم سمت راست این توزیع را در کنار دو توزیع نرمال و کوشی نشان می‌دهد. با توجه به شکل، کاملاً واضح است که دم توزیع HS بین دو توزیع نرمال (با دمی سبک) و کوشی (با دمی سنگین) قرار دارد.



شکل (۱): نمودار توابع چگالی توزیع‌های HS، نرمال و کوشی (سمت راست) و دم‌های سمت راست آن‌ها (سمت چپ)

۳- چارچوب مدل‌بندی GAMLSS

رده GAMLSSs یک چارچوب مدل‌بندی کلی و کاملاً منعطف برای یک متغیر پاسخ فراهم کرده است. یک طیف وسیع از توزیع‌های پیوسته و گسسته که توزیع‌های چوله و دم‌سنگین را نیز شامل می‌شوند، برای توزیع متغیر پاسخ معرفی و در بسته‌افزار *gamlss* در محیط R تعبیه شده‌اند. برای مشاهده فهرست این توزیع‌ها می‌توانید به استاسینوپولوس و همکاران [۸] مراجعه کنید. برخی از این توزیع‌ها تا چهار پارامتر را شامل می‌شوند که می‌توان آن‌ها را با μ، σ، ν و τ نشان داد که ه ترتیب بیانگر پارامترهای مکان (مثل میانگین)، مقیاس (مثل انحراف معیار) و شکل (مثل چولگی و کشیدگی) هستند. در چارچوب مدل‌بندی GAMLSS همه پارامترهای توزیع پاسخ را می‌توان به‌صورت پارامتری یا ترکیبی جمعی از توابع ناپارامتری هموار از متغیرهای

تبیینی مدل‌بندی کرد. البته می‌توان آمیخته‌ای از دو بخش پارامتری و ناپارامتری را به‌صورت مدلی نیمه‌پارامتری در نظر گرفت.

برای شروع فرض کنید مشاهدات y_i، برای $i = ۱,۲,\ldots,n$، به‌طور مستقل دارای تابع (چگالی) احتمال $f(y_i \mid \boldsymbol{\theta}^i)$ با بردار پارامترهای $\boldsymbol{\theta}^i = (\theta_{۱i}, \theta_{۲i}, \theta_{۳i}, \theta_{۴i})' = (\mu_i, \sigma_i, \nu_i, \tau_i)'$ باشند. بنا بر ریگبی و استاسینوپلوس [۳] چارچوب یک مدل GAMLSS به‌صورت زیر تعریف می‌شود.

فرض کنید تابع $g_k(\cdot)$، برای $k = ۱,۲,۳,۴$، یک تابع پیوند معلوم یکنوا باشد که پارامتر θ_k را به متغیرهای تبیینی مرتبط می‌کند. بنا؛ این می‌توان نوشت

$$g_k(\theta_k) = \boldsymbol{X}_k \boldsymbol{\beta}_k + \sum_{j=۱}^{J_k} h_{jk}(x_{jk}) \qquad (۱)$$

که در آن، برای هر k و $j = ۱,\cdots,J_k$، $\boldsymbol{\beta_k} = (\beta_{۱k}, \beta_{۲k}, \ldots, \beta_{J'_k k})'$ بردار پارامترهای J'_k بعدی، $\boldsymbol{X}_k$ ماتریس طرح معلوم $n \times J'_k$ بعدی، $h_{jk}(.)$ یک تابع ناپارامتری هموار از متغیر تبیینی X_{jk} و x_{jk}ها بردارهایی با طول n از مشاهدات متغیرهای تبیینی با اثرات غیرخطی هستند. مدل (۱) یک مدل نیمه‌پارامتری را در چارچوب مدل‌های GAM برای پارامتر θ_k تشکیل می‌دهد.

در مدل (۱) اثرات هموار متغیرها را می‌توان با کمک رهیافت بسط توابع پایه به‌صورت یک اثر تصادفی، مشابه $h(\boldsymbol{x}) = \boldsymbol{Z}\boldsymbol{\gamma}$، نمایش داد که در آن $\boldsymbol{Z}$ یک ماتریس پایه است که بر اساس مقادیر $\boldsymbol{x}$ ساخته می‌شود و $\boldsymbol{\gamma}$ برداری از ضرایب (اثرات) تصادفی است که فرض می‌شود دارای توزیع نرمال چندمتغیره است. با این نمایش، واضح است که می‌توان اثرات تصادفی را نیز به پیشگوی (۱)، برای هر پارامتر، به‌صورت جمعی اضافه کرد. بناب؛ ین مدل‌بندی داده‌های وابسته مثل داده‌های طولی، سری زمانی و فضایی در چارچوب رده مدل‌های GAMLSS نیز امکان‌پذیر است.

با این مقدمه، مدل (۱) را برای هرکدام از پارامترهای مکان، مقیاس و شکل می‌توان به‌صورت زیر نمایش داد:

$$g_۱(\boldsymbol{\mu}) = \boldsymbol{X}_۱\boldsymbol{\beta}_۱ + \sum_{j=۱}^{J_۱} \boldsymbol{Z}_{j۱}\boldsymbol{\gamma}_{j۱}$$

$$g_۲(\boldsymbol{\sigma}) = \boldsymbol{X}_۲\boldsymbol{\beta}_۲ + \sum_{j=۲}^{J_۲} \boldsymbol{Z}_{j۲}\boldsymbol{\gamma}_{j۲}$$

$$g_3(\boldsymbol{v}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=3}^{J_3}\boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=4}^{J_4}\boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}$$

که در آن $\boldsymbol{\mu}$، $\boldsymbol{\sigma}$، $\boldsymbol{v}$ و $\boldsymbol{\tau}$ بردارهای n بعدی هستند و $\boldsymbol{\beta_k}$ و $\boldsymbol{X}_k$ مشابه قبل تعریف می‌شوند. همچنین $\boldsymbol{Z}_{jk}$ ماتریس پایه معلوم $n\times q_{jk}$ بعدی وابسته به متغیرهای تبیینی هستند و $\boldsymbol{\gamma}_{jk}$ بردار تصادفی q_{jk} بعدی با پذیره $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}\left(0, \boldsymbol{G}_{jk}^{-1}\right)$ است که در آن $\boldsymbol{G}_{jk}^{-1}$ معکوس (تعمیم‌یافته) $q_{jk}\times q_{jk}$ بعدی ماتریس متقارن $\boldsymbol{G}_{jk} = \boldsymbol{G}_{jk}\left(\lambda_{jk}\right)$ است که ممکن است به ابرپارامتر λ_{jk} وابسته باشد. اگر ماتریس $\boldsymbol{G}_{jk}$ ناویژه باشد، بردار ضرایب $\boldsymbol{\gamma}_{jk}$ دارای یک توزیع ناسره است که چگالی آن متناسب است با $\exp(-\frac{1}{2}\lambda_{jk}\boldsymbol{\gamma}_{jk}'\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk})$. به‌عنوان‌مثال می‌توان به اثر تصادفی اتورگرسیو شرطی فضایی[1] (CAR، رو و هلد [۱۹]) برای داده‌های فضایی مشبکه‌ای اشاره کرد که یک توزیع ناسره دارد.

تعیین ساختارهای مختلف برای ماتریس‌های $\boldsymbol{Z}$ و $\boldsymbol{G}$ انواع مختلفی از اثرات جمعی را نتیجه می‌دهد. این اثرات شامل مؤلفه‌های تصادفی، مؤلفه‌های هموار متغیرها، مؤلفه‌های سری زمانی و اثرات فضایی می‌شوند. برای مؤلفه‌های هموار انتخاب‌های مختلفی مانند اسپلاین‌های تاوانیده[2] (ایلرز و مارکس [۲۰])، اسپلاین‌های درجه سوم[3]، چندجمله‌ای‌های موضعی وجود دارند.

چارچوب مدل‌بندی GAMLSS این امکان را فراهم می‌کند که توزیع شرطی متغیر پاسخ (به‌شرط اثرات تصادفی) هر توزیع دلخواهی چه از خانواده توزیع‌های نمایی چه خارج از آن باشد. این‌یک ویژگی خیلی منعطف و خوب محسوب می‌شود. البته به‌عنوان یک محدودیت، توزیع اثرات تصادفی باید نرمال باشد. در این مقاله، ما با تعبیه کردن توزیع HS به چارچوب رده مدل‌های GAMLSS امکان مدل‌بندی منعطف داده‌های پیوسته‌ای را فراهم کرده‌ایم که یک یا چند داده پرت را شامل می‌شوند. مدل موردنظر ما یک مدل نیمه‌پارامتری مکان-مقیاس است که در ساختار مدل (۱) با دو پارامتر $\theta_1 = \mu$ و $\theta_2 = \sigma$ صدق می‌کند.

۳-۱- استنباط مبتنی بر درستنمایی

رهیافت استنباطی مدنظر برای مدل پیشنهادی، مبتنی بر درستنمایی است. با توجه به این‌که در این مدل از مؤلفه‌های ناپارامتری جمعی مختلف می‌توان استفاده کرد، برای جلوگیری از پیچیدگی

1- Spatial Conditional Autoregressive
2- Penalized splines
3- Cubic splines

بیش‌ازحد، از یک چارچوب درستنمایی تاوانیده استفاده می‌شود. تحت مستقل بودن مشاهدات پاسخ، تابع لگاریتم درستنمایی مدل GAMLSS پیشنهادی به‌صورت

$$\ell(\boldsymbol{\mu},\boldsymbol{\sigma}) = \sum_{i=1}^{n} \log f\left(y_i;\mu_i,\sigma_i\right)$$

است که در آن

$$\ell(\mu,\sigma) = -\log\pi - \log\sigma + \log 2 + \frac{y-\mu}{\sigma} - \log\left(e^{2\frac{y-\mu}{\sigma}} + 1\right)$$

و پارامترهای مکان و مقیاس توزیع HS با تابع چگالی $f(.)$ توسط مدل (۱) به متغیرهای تبیینی مرتبط می‌شوند. تابع درستنمایی تاوانیده برای این مدل به‌صورت زیر است:

$$\ell_p = \ell(\boldsymbol{\mu},\boldsymbol{\sigma}) - \frac{1}{2}\sum_{k=1}^{2}\sum_{j=1}^{J_k}\boldsymbol{\gamma}_{kj}^T \boldsymbol{G}_{kj}\left(\lambda_{kj}\right)\boldsymbol{\gamma}_{kj}. \qquad (۲)$$

مجموعه پارامترهایی که باید برآورد شوند، عبارتند از بردار پارامترهای مؤلفه خطی مدل، یعنی $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right)'$، بردار پارامترهای اثرات تصادفی (توابع ناپارامتری)، یعنی

$$\boldsymbol{\gamma} = \left(\gamma_{11},\cdots,\gamma_{\mathbf{J}_11},\gamma_{12}\cdots,\gamma_{\mathbf{J}_22}\right)'$$

و ابرپارامترهای مدل، یعنی $\boldsymbol{\lambda} = \left(\lambda_{11},\cdots,\lambda_{\mathbf{J}_11},\lambda_{12}\cdots,\lambda_{\mathbf{J}_22}\right)'$. در چارچوب مدل‌های GAMLSS، پارامترهای خطی $\boldsymbol{\beta}$ و پارامترهای اثرات تصادفی $\boldsymbol{\gamma}$، به ازای مقادیر ثابتی برای ابرپارامترهای هموارساز $\boldsymbol{\lambda}$، با ماکسیمم کردن تابع درستنمایی تاوانیده ℓ_p در (۲) برآورد می‌شوند. برای این کار دو روش پایه‌ای با نام‌های الگوریتم‌های RS[1] و CG[2] (ریگبی و استاسینوپلوس [۳]) وجود دارند. هر دوی این روش‌ها از یک الگوریتم کمترین توان‌های دوم (تاوانیده) موزون تکراری برای محاسبه برآوردهای ML پارامترهای مذکور، با مفروض بودن ابرپارامترهای هموارساز، استفاده می‌کنند. ریگبی و استاسینوپلوس [۳] نشان دادند برآوردهای حاصل از این روش‌ها، با در نظر گرفتن توزیع‌های پیشین ناآگاهی‌بخش تخت برای ضرایب رگرسیونی و نرمال برای ضرایب تصادفی $\boldsymbol{\gamma}$، برآوردهای مد پسینی[3] (MAP) در چارچوب یک رهیافت بیز تجربی نیز هستند.

1- Rigby and Stasinopoulos
2- Cole and Green
3- Maximum a posteriori

روش RS تعمیم الگوریتم به کار گرفته‌شده توسط ریگبی و استاسینوپلوس [۱۰] است. روش CG نیز تعمیم الگوریتم کول و گرین [۲۱] است که در آن مشتقات مرتبه اول تابع درستنمایی و مقادیر امید ریاضی مشتقات مرتبه دوم تابع درستنمایی نسبت به پارامترهای توزیع یعنی μ و σ موردنیاز هستند. البته برای مشتقات مرتبه دوم، مشابه روش بهینه‌سازی FGS[1] (برویدن [۲۲])، می‌توان از تقریب‌های عددی نیز استفاده کرد. مشتقات اول توزیع HS نسبت به μ و σ عبارتند از

$$\frac{\partial \ell(\mu,\sigma)}{\partial \mu} = -\frac{۱}{\sigma} + \frac{۲}{\sigma}\frac{exp(\frac{۲(y-\mu)}{\sigma})}{exp(\frac{۲(y-\mu)}{\sigma})+۱}$$

$$\frac{\partial \ell(\mu,\sigma)}{\partial \sigma} = -\frac{۱}{\sigma} - \frac{y-\mu}{\sigma^{۲}} + ۲\frac{y-\mu}{\sigma^{۲}}\frac{exp(\frac{۲(y-\mu)}{\sigma})}{exp(\frac{۲(y-\mu)}{\sigma})+۱}.$$

برای مشتقات دوم نیز از تقریب‌های عددی استفاده کرده‌ایم. استاسینوپولوس و ریگبی [۹] تأکید کردند که الگوریتم CG نسبت به الگوریتم RS کلی‌تر ولی در همگرایی کندتر است. آن‌ها نشان دادند در مواردی که پارامترهای توزیع از نظر اطلاع متعامد هستند (به این معنی که امید ریاضی مشتقات حاصل‌ضرب تابع درستنمایی نسبت به پارامترها صفر هستند)، استفاده از روش RS به دلیل سرعت بالاتر آن توصیه می‌شود؛ و در مواردی که پارامترها متعامد نباشند، الگوریتم CG از کارایی بیشتری برخوردار است. البته یک روش سازوار[2] توسط استاسینوپولوس و ریگبی [۹] پیشنهاد شده است که از ترکیب هر دو الگوریتم در بهینه‌سازی تابع درستنمایی تاوانیده استفاده می‌کند.

برای برآورد ابرپارامترهای λ دو روش موضعی[3] (ریگبی و استاسینوپولوس [۲۳]) و فراموضعی[4] (ریگبی و استاسینوپولوس [۳]) پیشنهاد شده‌اند. روش موضعی سریع‌تر و اجرای آن ساده‌تر است. در واقع در روش موضعی برآورد ابرپارامترها در هر مرحله الگوریتم‌های RS یا CG انجام می‌شود، درحالی‌که در روش فراموضعی، حداقل سه معیار برای برآورد ابرپارامترها وجود دارند:

1- Broyden-Fletcher-Goldfarb-Shanno
2- Adaptive
3- Locally
4- Globally

۱. مینیمم کردن معیار اعتبارسنجی متقابل تعمیم‌یافته[1] (GCV؛ وود [۲۴])

۲. مینیمم کردن معیار اطلاع آکاییک (AIC، آکاییک [۲۵]) یا معیار اطلاع بیزی شوارتز (BIC؛ شوارتز [۲۶])

۳. مقدار ماکسیمم درستنمایی.

ما در این مقاله، از معیار سوم و به‌صورت موضعی استفاده کرده‌ایم.

برای مقایسه و انتخاب مدل برتر در بین مدل‌های نامزد می‌توان از معیارهای AIC و BIC استفاده کرد. بررسی نیکویی برازش مدل‌ها نیز توسط تحلیل باقی‌مانده‌ها انجام می‌شود.

۴- مطالعه شبیه‌سازی

برای ارزیابی عملکرد مدل پیشنهادی، در این بخش از یک مثال شبیه‌سازی استفاده کردیم. هدف از ارزیابی، مقایسه عملکرد مدل پیشنهادی با دو مدل نرمال (به‌عنوان یک مدل با دم سبک) و مدل رگرسیونی t-استودنت (به‌عنوان یک مدل با دم سنگین) بود. دلایل مقایسه با این دو مدل عبارتند از

- مدل مکان-مقیاس جدید HS را به‌عنوان جانشینی نیرومند[2] برای مدل مکان-مقیاس نرمال معرفی کردیم. برای همین برازش هر دو مدل HS و نرمال موردنظر بودند.
- مدل مکان-مقیاس مبتنی بر توزیع t-استودنت نیز جانشینی نیرومند برای مدل نرمال محسوب می‌شود؛ اما با توجه به دم سنگین بودن این توزیع مایل بودیم نتیجه استفاده از آن و مدل با دم نیمه‌سنگین HS را بدانیم.

بنا به پیشنهاد یکی از داوران، برای شبیه‌سازی داده‌ها از هیچ‌کدام از مدل‌های رقیب استفاده نشد. داده‌ها را از یک مدل رگرسیونی مکان-مقیاس بر پایه توزیع لجستیک تولید کردیم. توزیع لجستیک دمی سنگین‌تر از نرمال دارد ولی به‌اندازه توزیع t-استودنت به‌عنوان یک توزیع دم سنگین معروف و معمول نیست. برای تولید داده‌ها فرض کردیم مشاهده i ام متغیر پاسخ دارای توزیع لجستیک با پارامتر مکان μ_i و پارامتر مقیاس σ_i باشد، به‌طوری‌که

$$\mu_i = x_{1i} + \sin(\frac{\pi}{5} x_{2i})$$

و

1- Generalized Cross Validation (GCV)
2- Robust

$$\sigma_i = \exp(x_{3i} + \sin(\frac{\pi}{5} x_{4i})) \qquad (۳)$$

که در آن، برای $i = 1,2,\ldots,n$، همه متغیرهای تبیینی x_1, x_2, x_3, x_4 از توزیع نرمال با میانگین صفر تولید شدند. انحراف معیار دو متغیر x_1, x_3 برابر ۰/۲ و دو متغیر تبیینی دیگر برابر ۲ انتخاب شدند. همان‌طور که معلوم است در هر دو پارامتر مکان و مقیاس توزیع، هم اثر خطی وجود دارد و هم مؤلفه غیرخطی. حجم نمونه را برابر ۱۰۰، ۲۰۰ و ۵۰۰ در نظر گرفتیم و برای هرکدام از آن‌ها ۲۰۰ مجموعه داده را از مدل (۳) تولید کردیم. برای برازش مدل از روش‌های تشریح شده در بخش ۳-۱ استفاده کردیم. برای مدل‌بندی مؤلفه ناپارامتری جمعی در هر دو پارامتر نیز از اسپلاین‌های تاوانیده استفاده کردیم.

برای برازش مدل‌های نرمال و t-استودنت از بسته *gamlss* در نرم‌افزار R استفاده کردیم که هر دو توزیع را در خود دارد. ولی مدل HS در این بسته وجود ندارد. پدیدآوران بسته *gamlss* این امکان را فراهم کرده‌اند که بتوان با کدنویسی در محیط نرم‌افزار R از امکانات مدل‌بندی و برازش موجود در این بسته برای سایر توزیع‌های جدید توسط سایرین استفاده کرد. با این امکان، از قابلیت موجود برای معرفی یک مدل نیمه‌پارامتری بر پایه توزیع HS بهره برده و کدهای مورداشاره همراه با مثال ارائه‌شده در مقاله را در لینک زیر[1] در دسترس خوانندگان قرار داده‌ایم.

برای ارزیابی برآوردهای پارامترهای خطی از معیار میانگین توان‌های دوم خطا[2] (*MSE*) استفاده کردیم که به‌صورت زیر تعریف می‌شود:

$$MSE\left(\hat{\beta}_j\right) = \frac{1}{200}\sum_{j=1}^{200}\left(\hat{\beta}_j - \beta\right)^2$$

که در آن $\hat{\beta}_j$ ضریب رگرسیونی برآوردشده در مجموعه داده تولیدشده j ام و β مقدار واقعی آن است. برای ارزیابی مدل‌بندی ناپارامتری و برآورد مؤلفه غیرخطی در هر دو پارامتر مکان و مقیاس از معیار MSE تجمیع‌شده[3] (*IMSE*؛ ساکس و همکاران [۲۷]) با تعریف زیر استفاده کردیم:

$$IMSE\left(\hat{f}_j\right) = \frac{1}{200}\sum_{j=1}^{200}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_j\left(x_i\right) - f\left(x_i\right)\right)^2$$

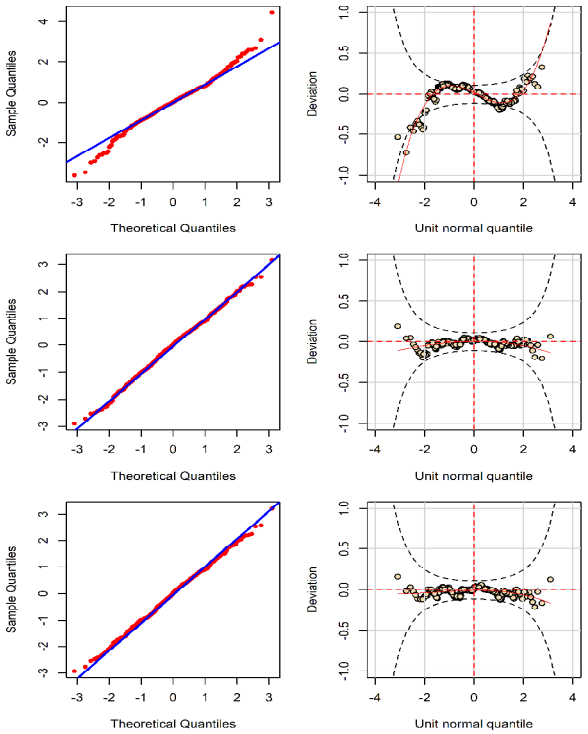
1- https://github.com/JamilOwnuk/GAMLSS-HS/blob/master/HS%20distribution
2- Mean Squared Errors
3- Integrated Mean Squared Errors

که در آن $\hat{f}_j(x)$ تابع برآوردشده برای مجموعه داده j ام و x_i ها مقادیری معلوم هستند که برای محاسبه اختلاف تابع برآوردشده و تابع واقعی در نظر گرفته می‌شوند؛ در این مثال مقادیر تولیدشده متغیرهای تبیینی را برای مقداردهی به کار گرفتیم.

نتایج دو معیار برای هر سه مدل رقیب در جدول (۱) گزارش شده‌اند. با توجه به نتایج جدول، با افزایش حجم نمونه مقادیر MSE و IMSE اثرات برآوردشده کاهش پیدا می‌کنند؛ بنابراین می‌توان برقراری نتایج نظری سازگاری مجانبی در این مدل‌ها را نتیجه گرفت. از طرفی، عملکرد دو مدل HS و t-استودنت در برآورد پارامترهای رگرسیونی و همین‌طور تابع جمعی هم برای مکان و هم مقیاس برتر از مدل نرمال است؛ اما برتری مشخصی برای دو مدل HS نسبت به t-استودنت یا برعکس در پارامتر مکان دیده نمی‌شود. در مقابل، برای پارامتر مقیاس مدل رگرسیونی، هم در برآورد ضرایب خطی و هم تابع جمعی مدل HS کاملاً نسبت به مدل t-استودنت برتر است؛ بنابراین می‌توان در مجموع برتری مدل HS را در برازش داده‌ها نتیجه گرفت.

جدول (۱): مقادیر MSE پارامترهای خطی و IMSE اثر جمعی برای داده‌های شبیه‌سازی‌شده

حجم نمونه	مدل	پارامتر مکان			پارامتر مقیاس		
		جمعی	عرض از مبدأ	شیب	جمعی	عرض از مبدأ	شیب
۱۰۰	نرمال	۰/۰۵۶۱	۰/۰۳۰۴	۰/۴۶۵۳	۰/۳۴۴۹	۰/۳۰۶۳	۰/۲۸۴۷
	T	۰/۰۵۷۴	۰/۰۳۱۱	۰/۴۴۳۳	۰/۰۵۶۵	۰/۱۸۹۵	۰/۲۷۷۲
	HS	۰/۰۵۴۱	۰/۰۲۹۵	۰/۴۶۲۵	۰/۰۵۳۹	۰/۰۲۸۷	۰/۲۷۹۲
۲۰۰	نرمال	۰/۰۳۴۷	۰/۰۱۸۵	۰/۱۹۶۵	۰/۳۴۶۲	۰/۳۳۴۰	۰/۱۱۰۹
	t	۰/۰۳۴۶	۰/۰۱۹۷	۰/۱۸۱۴	۰/۰۴۹۹	۰/۲۰۲۷	۰/۱۰۷۵
	HS	۰/۰۳۲۳	۰/۰۱۶۸	۰/۱۸۴۹	۰/۰۴۴۰	۰/۰۳۷۲	۰/۱۰۳۷
۵۰۰	نرمال	۰/۰۲۱۸	۰/۰۱۱۳	۰/۰۷۰۵	۰/۳۵۹۴	۰/۳۷۳۳	۰/۰۴۹۶
	t	۰/۰۲۴۴	۰/۰۱۰۹	۰/۰۶۳۴	۰/۰۶۰۰	۰/۲۲۲۱	۰/۰۴۵۲
	HS	۰/۰۲۸۵	۰/۰۱۲۱	۰/۰۷۳۲	۰/۰۴۳۶	۰/۰۵۵۰	۰/۰۴۱۷



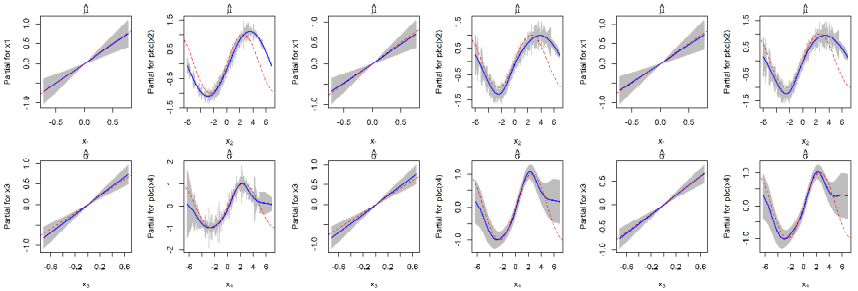
شکل (۲): نمودارهای چندک-چندک (چپ) و کرمی (راست) باقی‌مانده‌های حاصل از سه مدل نرمال (بالا)، t-استودنت (وسط) و HS (پایین) برای یک مجموعه داده شبیه‌سازی‌شده با حجم نمونه ۵۰۰

برای یک مجموعه از داده‌های شبیه‌سازی‌شده با حجم نمونه ۵۰۰ نمودارهای چندک-چندک[1] باقی‌مانده‌های حاصل از هر سه مدل در شکل (۲) نمایش داده شده‌اند. نمودارهای سمت راست در شکل (۲) به نمودار کرمی[2] معروف هستند که اولین بار توسط ون بورن و فردریکس (۲۰۰۱) معرفی شدند. نمودار کرمی نوعی نمودار چندک-چندک باقی‌مانده‌های روندزدایی‌شده[3] است که نشان‌دهنده فاصله توزیع متغیر پاسخ مشاهده‌شده از توزیع مفروض، به‌عنوان توزیع واقعی پاسخ، است. در صورت نیکویی برازش توزیع مفروض به داده‌ها، نقاط روندزدایی‌شده می‌توانند در یک کران محدودشده پراکنش داشته باشند. انحنای پراکنش نقاط به‌صورت درجه دو یا درجه سه و همچنین خارج از کران قرار گرفتن نقاط، نشان از عدم‌کفایت مدل برازش‌شده دارد. برای توضیح

1- Quantile-Quantile plot
2- Worm plot
3- Quantile-Quantile plot for detrended residuals

بیشتر در مورد تفسیر این نوع نمودار می‌توانید منبع ون بورن و فردریکس [۲۸] را ببینید. با توجه به شکل (۲) واضح است که مدل مکان-مقیاس نرمال برای این داده‌ها برازش نامناسبی دارند و در مقابل هر دو مدل HS و t-استودنت برازشی کاملاً مناسب و قابل‌قبول دارند.

شکل (۳) نیز اثرات برآوردشده خطی و غیرخطی را برای هر دو پارامتر مکان و مقیاس در سه مدل نرمال، t-استودنت و HS نشان می‌دهد. به همراه برآورد اثرات، کران‌های اطمینان ۹۵ درصد نیز رسم شده‌اند که بر اساس توزیع مجانبی نرمال برآوردگرهای ML محاسبه شده‌اند (ریگبی و استاسینوپولوس [۳]). در برآورد ضرایب خطی، هر سه مدل تقریباً یکسان عمل کرده‌اند؛ اما با توجه به کران‌های اطمینان محاسبه‌شده، مدل HS نسبت به دو مدل دیگر در برآورد اثرات غیرخطی کاراتر عمل کرده است.

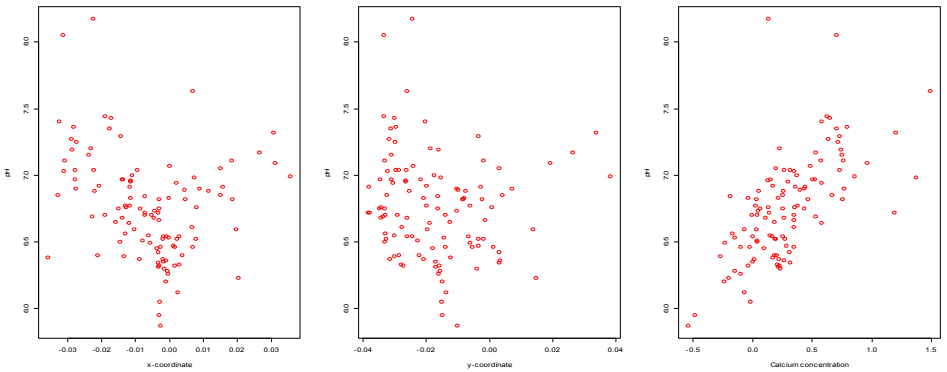


شکل (۳): اثرات خطی و غیرخطی برآوردشده (ممتد) و واقعی (خط‌چین) پارامترهای مکان و مقیاس مدل‌های نرمال (سمت راست)، t-استودنت (وسط) و HS (سمت چپ) برای حجم نمونه ۵۰۰

۵- مثال واقعی

برای نمایش کاربست و برتری ممکن مدل پیشنهادی جدید، تحلیل رگرسیونی یک مجموعه داده واقعی را که در داگلاس و دلامپادی [۱۱] گزارش شده است، مدنظر قرار دادیم. هدف از تحلیل این داده‌ها بررسی اسیدی بودن دریاچه‌های واقع در رشته‌کوه تیغه آبی در کشور آمریکا است. این مجموعه داده از ۱۱۲ مشاهده برای ۴ متغیر تشکیل شده است. متغیر پاسخ مقدار pH دریاچه‌ها (y) و متغیرهای تبیینی رگرسیونی مختصات مرکزی هر دریاچه، شامل طول جغرافیایی (x_1) و عرض جغرافیایی (x_2) و گاریتم غلظت کلسیم با واحد میلی‌گرم در هر لیتر است (x_3) که گاریتم آن در مبنای ۱۰ محاسبه شده است. این مجموعه داده در بسته‌افزار *assist* در نرم‌افزار R در دسترس است.

شکل (۴) نمودارهای پراکنش کناری متغیر پاسخ در مقابل متغیرهای تبیینی را نشان می‌دهد. از روی این نمودارها چند نکته قابل بیان هستند:



شکل (۴): نمودارهای پراکنش کناری پاسخ و متغیرهای تبیینی برای داده‌های دریاچه‌های اسیدی

۱. نمودار پراکنش پاسخ در مقابل لگاریتم غلظت کلسیم (نمودار سمت راست) یک رابطه مثبت را به همراه نوسان بیشتر برای مقادیر بزرگ‌تر لگاریتم غلظت نشان می‌دهد. این شهود نشانی از برقرار نبودن پذیره هم‌واریانسی در مدل رگرسیون مکانی است.

۲. نمودار پراکنش پاسخ در مقابل طول جغرافیایی، یک روند غیرخطی کاهشی و سپس افزایشی را به همراه نوسان ناهمگن در طی این روند نمایش می‌دهد. در مقابل روند مشخص و قابل درکی در مقابل متغیر عرض جغرافیایی ملموس نیست.

۳. جدا از ناهم‌واریانسی که در مشاهدات مشهود است، دو داده پرت در هر سه نمودار پراکنش که کاملاً از سایر مشاهدات جدا هستند، دیده می‌شوند. این مسئله بر لزوم لحاظ کردن آن‌ها توسط مدلی که توانایی برخورد با داده‌های پرت را (که تعداد اندکی از کل مشاهدات را شامل می‌شوند) دارند، تأکید دارد.

این سه نکته، حاکی از پیچیدگی مدل‌بندی این داده‌ها و کافی نبودن مدل میانگین است. برای مدل‌بندی داده‌ها دو مدل میانگین (فقط مکان) (مدل ۴) و مکان-مقیاس (مدل ۵) را با تنها دو متغیر تبیینی X_1 و X_2 به‌صورت زیر در نظر گرفتیم:

$$\begin{cases} \mu_i = \beta_0 + h_1(x_1) + h_2(x_2) \\ \log\sigma_i = \beta_1 \end{cases} \qquad (۴)$$

و

$$\begin{cases} \mu_i = \beta_2 + h_1(x_1) + h_2(x_2) \\ \log\sigma_i = \gamma_0 + \gamma_1 x_1 + h_2(x_2) \end{cases} \qquad (۵)$$

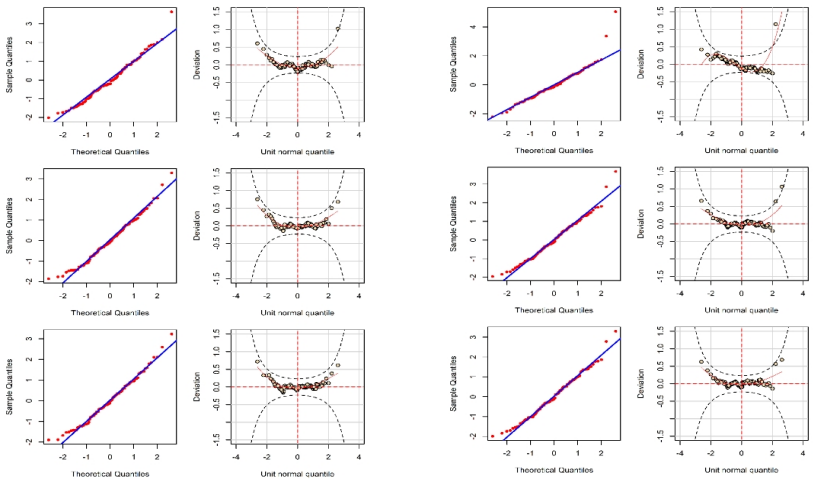
در اینجا نیز، با توجه به حضور دو داده پرت، برای مقایسه مدل پیشنهادی ما با مدلی که قادر به مدل‌بندی داده‌های پرت باشد، از مدل مکان و مکان-مقیاس t-استودنت در کنار مدل نرمال استفاده کردیم. نتایج برازش هر سه مدل در جدول (۲) گزارش شده‌اند. از نظر معیارهای AIC و BIC می‌توان گفت مدل مکان (۴) نسبت به مدل مکان-مقیاس (۵) ضعیف‌تر است. در هر دو نوع مدل‌بندی نیز مدل پیشنهادی حتی نسبت به مدل نیرومند t-استودنت برتر است و می‌توان نتیجه گرفت که در حضور داده‌های با دم نیمه‌سنگین مدل‌های دم‌سنگینی مثل t-استودنت به‌خوبی مدل پیشنهادی نخواهند بود.

نمودارهای نیکویی برازش مبتنی بر باقی‌مانده‌ها برای هر سه مدل نیز در شکل (۵) رسم شده‌اند که نشان می‌دهند مدل نرمال از دو مدل دیگر از کارایی کمتری برخوردار است. نیکویی برازش مدل‌های HS و t-استودنت به‌ویژه برای مدل (۵) قابل تائید است.

جدول (۲): نتایج برازش مدل‌های (۴) و (۵) بر روی داده‌های دریاچه‌های اسیدی

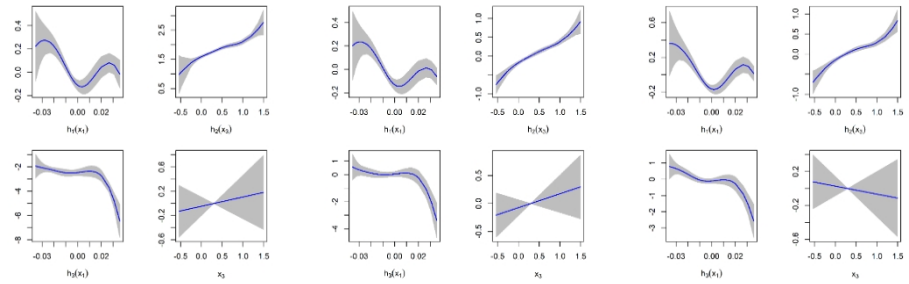
مدل	توزیع	β_0	β_1	β_2	γ_0	γ_1	ν	AIC	BIC
مدل (۴)	نرمال	۶/۵۰۵	-۱/۴۰۴	-	-	-	-	۲۳/۲۷	۵۰/۴۶
	T	۶/۴۹۴	-۱/۶۹۵	-	-	-	۱/۵۸۱	۸/۳۰	۳۸/۲۱
	HS	۶/۵۵۲	-۱/۸۹۱	-	-	-	-	۷/۰۴	۳۴/۲۳
مدل (۵)	نرمال	-	-	۶/۴۹۹	-۱/۶۳۳	-۲۴/۶۵۰	-	۹/۱۳	۴۹/۹۱
	T	-	-	۶/۴۷۸	-۱/۸۹۸	-۱۹/۵۳۸	۱/۶۶۱	۱۱/۳۵	۵۴/۸۵
	HS	-	-	۶/۳۸۶	-۱/۸۰۵	۳/۳۷۴	-	۶/۹۴	۴۷/۷۲

منحنی‌های اثرات برآوردشده خطی و غیرخطی برای هر دو پارامتر مکان و مقیاس تحت سه توزیع نرمال، t-استودنت و HS برای مدل (۵) در شکل (۶) رسم شده‌اند. اثر برآوردشده برای متغیر طول جغرافیایی در مقیاس مدل نرمال کاملاً در تناقض با دو مدل دیگر است؛ در واقع شیب اثر در مدل نرمال منفی ولی در دو مدل دیگر مثبت است. این اتفاق منجر به یک نتیجه‌گیری نادرست خواهد شد. سایر اثرها در هر سه مدل برای هر دو پارامتر مکان و مقیاس تقریباً شبیه به هم برآورد شده‌اند، با این تفاوت که عدم قطعیت برآوردشده برای متغیر تبیینی لگاریتم غلظت کلسیم در کران‌های مشاهدات در پارامتر مکان مدل HS از دو مدل رقیب بیشتر است.



(ب) (الف)

شکل (۵): نمودارهای چندک-چندک و کرمی مدل‌های نرمال (بالا)، HS (وسط) و t-استودنت (پایین) برای مدل‌های (۴) (الف) و (۵) (ب)



(ج) (ب) (الف)

شکل (۶): منحنی‌های اثرات برآوردشده پارامترهای مکان (بالا) و مقیاس (پایین) در مدل‌های نرمال (الف)، HS (ب) و t-استودنت (ج) برای مدل (۵)

۶- نتیجه‌گیری

پیچیدگی‌های ذاتی موجود در داده‌های عملی، محققان را بر آن داشته است تا به دنبال معرفی و بسط مدل‌های منعطفی برای مدل‌بندی واقعی آن‌ها باشند. رده مدل‌های GAMLSS یکی از تلاش‌های موفق در این زمینه است که در سال‌های اخیر طرفداران و کاربردهای زیادی پیدا کرده

است. با توجه به این واقعیت، در این مقاله یک مدل تنومند مکان-مقیاس، بر اساس توزیع هایپربولیک سکانت با دم نیمه‌سنگین، معرفی کردیم که از توانایی بالاتری نسبت به مدل‌های تنومند دم‌سنگین در برخورد با داده‌های پرت اندک نسبت به کل داده‌ها برخوردار است. این مدل جدید عضوی از رده مدل‌های GAMLSS محسوب می‌شود و بنابراین از مزایای استنباطی توسعه‌یافته برای این مدل‌ها برخوردار است.

برای برازش مثال‌های شبیه‌سازی و کاربردی، در بخش مدل‌بندی مؤلفه‌های غیرخطی، از اسپلاین‌های تاوانیده استفاده کردیم. جانشین‌های مختلف دیگری مانند اسپلاین‌های درجه سوم و چندجمله‌ای‌های موضعی برای این کار وجود دارند که می‌توان از قابلیت‌های آن‌ها نیز بهره برد. در هر دو مثال‌های شبیه‌سازی و کاربردی، عملکرد برتر مدل پیشنهادی ما نسبت به مدل‌های رقیب نمایش داده شد. این نتیجه می‌تواند در موقعیت‌های کاربردی حائز اهمیت باشد و مدل پیشنهادی HS را به‌عنوان یک امکان برای استفاده در جعبه‌ابزار آماردانان قرار دهد.

منابع

[1] Nelder, J. and Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society: Series A*, **135**, 370-384.

[2] Hastie, T. and Tibshirani, R. (1990). Generalized additive models, *Statistical Science*, **1**, 297-310.

[3] Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507-554.

[4] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American statistical Association*, **88**, 9-25.

[5] Lee, Y. and Nelder, J. (2001). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions, *Biometrika*, **88**, 987-1006.

[6] Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 381-400.

[7] Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **50**, 201-220.

[8] Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V. and De Bastiani, F. (2017). *Flexible Regression and Smoothing using GAMLSS in R*, Chapman and Hall/CRC, London.

[9] Stasinopoulos, M.D. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R, *Journal of Statistical Software*, **23**, 1-46.

[10] Rigby, R.A. and Stasinopoulos, D.M. (1996). A semi-parametric additive model for variance heterogeneity, *Statistics and Computing*, **6**, 57-65.

[11] Douglas, A. and Delampady, M. (1990). Eastern lake survey–phase I: documentation for the data base and the derived data sets, *SIMS Technical Report*, 160.

[12] Fisher, R.A. (1921). 014: On the "Probable Error" of a coefficient of correlation deduced from a small Sample, *Metron*, **1**, 3-32.

[13] Dodd, E.L. (1925). The frequency law of a function of variables with given frequency laws, *Annals of Mathematics*, **27**, 12-20.

[14] Perks, W. (1932). On some experiments in the graduation of mortality statistics, *Journal of the Institute of Actuaries*, **63**, 12-57.

[15] Vaughan, D.C. (2002). The generalized secant hyperbolic distribution and its properties, *Communications in Statistics-Theory and Methods*, **31**, 219-238.

[16] M.J. Fischer, and D. Vaughan, Classes of skew generalized hyperbolic secant distributions, Diskussionspapiere//Friedrich-Alexander-Universitt Erlangen-Nrnberg, Lehrstuhl fr Statistik und konometrie, 2002.

[17] Fischer, M.J. (2013). *Generalized Hyperbolic Secant Distributions: With Applications to Finance*, Springer, New York.

[18] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, New York.

[19] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall/CRC, London.

[20] Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with b-splines and penalties, *Statistical Science*, **11**, 89-121.

[21] Cole, T.J. and Green, P.J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood, *Statistics in Medicine*, **11**, 1305-1319.

[22] Broyden, C.G. (1976). Quasi-Newton methods and their application to function minimization, *Mathematics of Computation*, **21**, 368-381.

[23] Rigby, R.A., Stasinopoulos, M.D. and Voudouris, V. (2013). Discussion: A comparison of GAMLSS with quantile regression, *Statistical Modelling*, **13**, 335-348.

[24] Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall, New York.

[25] Akaike, H. (1974). *A new look at the statistical model identification, Selected Papers of Hirotugu Akaike*, Springer, New York.

[26] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461-464.

[27] Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, **4**, 409-423.

[28] Buuren, S. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves, *Statistics in Medicine*, **20**, 1259-1277.

JAMM,
DOI: 10.22055/JAMM.2020.31201.1763
Shahid Chamran University of Ahvaz
ISSN: 2251-8088
Jamm.Scu.ac.ir

Semiparametric Location-Scale Regression Model With Semi-Heavy Tails Based On Hyperbolic Secant Distribution

Jamil Ownuk, Hossein Baghishani and Ahmad Nezakati

Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran

Received: September 26 2019 Accepted for publication: May 23 2020
Corresponding author: hbaghishani@shahroodut.ac.ir

Abstract

Practitioners who use the classical regression model have been realized that many of its assumptions seldom hold. We then need flexible models to capture the real intrinsic properties of data. The class of generalized additive models for location, scale, and shape is very flexible in analyzing the inherent complexity of the data. This class of models provides the ability to do regression modelling beyond the mean of the response variable. Indeed, to admit outliers in the modelling framework is vital. Where we have a few outliers, the model could be too complicated by using heavy-tailed distributions. To overcome this issue, in this paper, we introduce a new location-scale semiparametric regression that is constructed based on a semi-heavy-tailed distribution, named hyperbolic secant, in the considered class of the models. We explore the performance of the proposed model by a simulation study and compare the results with a classical normal model. We also illustrate the model in a real application.

***Keywords*:** Semi-heavy-tailed distribution, Hyperbolic secant distribution, Outlier, Penalized likelihood, Location-scale regression.

Mathematics Subject Classification (2010): 62J05, 62G08.