خانواده توزیع آمیخته-مقیاس چوله-نرمال و کاربرد آن در مدل‌های رگرسیونی غیرخطی بیزی

مهسا عابدینی[۱] ، ایرج کاظمی[۲]

چکیده

در بسیاری از تحقیقات قبلی برازش مدل‌های رگرسیونی غیرخطی نرمال به منظور تحلیل داده‌ها با ساختار پخش متقارن به صورت توابع غیرخطی از پارامترهای مجهول به کار رفته است. اما در عمل ممکن است توزیع مانده‌ها نامتقارن بوده و انتخاب توزیع نرمال مناسب نباشد. یک خانواده از توزیع‌های آماری که اخیراً مورد توجه قرار گرفته است آمیخته-مقیاس چوله-نرمال است که توزیع‌های چوله و دم-سنگینی مانند چوله-تی و چوله-اسلش را به عنوان حالات خاصی در بر می‌گیرد. از آن‌جایی که استنباط آماری پارامترها توسط روش حداکثر درستنمایی حاشیه‌ای منجر به حل انتگرال‌های پیچیده با ابعاد بالا خواهد شد، ما در این مقاله رهیافت شبیه‌سازی مونت کارلوی زنجیر مارکفی را برای استنباط بیزی پارامترهای مدل به کار می‌بریم. همچنین، مدل غیرخطی را با توزیع‌های پیشنهادی بر مجموعه‌ای از داده‌های واقعی برازش می‌دهیم تا اهمیت مدل پیشنهادی را بیان کنیم.

واژه‌های کلیدی: آمیخته-مقیاس، استنباط بیزی، رهیافت نمونه‌گیر گیبز، دم-سنگین، الگوریتم متروپلیس-هستینگز.

۱ مقدمه

در دهه اخیر تحقیق‌های متعددی جهت معرفی توزیع‌های آمیخته-مقیاس نرمال به عنوان جایگزینی مناسب برای نرمال انجام شده است که از آن جمله می‌توان به شکل آبادی و کاظمی (۱۳۹۱) اشاره نمود. این توزیع‌ها در تحلیل مشاهده‌ها با ساختار پخش متقارن و دم سنگین به عنوان جایگزین نرمال به کار می‌روند. در این مقاله، کاربرد خانواده جدیدی از توزیع‌های آمیخته-مقیاس چوله-نرمال ($SMSN$) را در برازش مدل‌های رگرسیونی غیرخطی بررسی می‌کنیم. برازش این مدل‌ها با توزیع‌های $SMSN$ و با به کارگیری الگوریتم EM توسط گری و همکاران (۲۰۱۱) انجام شده است. آنان همچنین استفاده از روش‌های بیزی مانند مونت کارلوی زنجیر مارکفی ($McMC$) را در برازش مدل‌های فوق پیشنهاد داده‌اند. استفاده از رهیافت $McMC$ و کاربردی از خانواده $SMSN$ در مدل‌های غیرخطی توسط کانچو و همکاران (۲۰۱۱) ارائه شده است. آن‌ها با استفاده از روش‌های بیزی بر اساس معیار کولبک-لیبر به تشخیص داده‌های پرنفوذ پرداختند.

در این مقاله تعدادی از اعضای خانواده $SMSN$ را در برازش مدل‌های رگرسیون غیرخطی در نظر گرفته و توسط روش‌های بیزی پارامترهای مدل را برآورد می‌کنیم. ساختار این مقاله به این صورت است که ابتدا در بخش دوم خانواده توزیع آمیخته-مقیاس چوله-نرمال معرفی می‌شود. آنگاه در بخش سوم مدل رگرسیونی غیرخطی را همراه با برخی از ویژگی‌های خاص آن معرفی می‌نماییم. در بخش چهارم مدل‌های رگرسیونی غیرخطی با خطای آمیخته-مقیاس چوله-نرمال ($SMSN-NLM$) را مورد بررسی قرار می‌دهیم. همچنین، توزیع‌های پسین شرطی کامل مورد نیاز جهت برآورد پارامترها با روش الگوریتم نمونه‌گیری گیبز را به‌دست می‌آوریم. در بخش پنجم با استفاده از روش شبیه‌سازی مونت کارلوی زنجیر مارکفی، مدل رگرسیونی غیرخطی با مانده‌های $SMSN$ را روی داده‌های واقعی برازش داده و با معیارهای انتخاب مدل، برازنده‌ترین مدل را انتخاب می‌نماییم.

[۱]دانشجوی کارشناسی ارشد دانشگاه اصفهان

[۲]عضو هیأت علمی گروه آمار دانشگاه اصفهان

۲ خانواده توزیع آمیخته–مقیاس چوله–نرمال

تعریف ۱.۲. متغیرهای تصادفی مستقل T_0 و T_1 دارای توزیع نرمال استاندارد را در نظر بگیرید. فرض کنید $k(u)$ تابعی مثبت از u باشد که در آن U متغیری تصادفی با تابع توزیع $H(.;v)$ است که $\mathbf{v}$برداری از پارامترها است. یک متغیر تصادفی، مانند Y، متعلق به خانواده توزیع $SMSN$ با نماد $SMSN(\mu,\sigma^2,\lambda,H)$ است هرگاه نمایش تصادفی آن به صورت

$$Y \overset{d}{=} \mu + \sigma(\delta T + k^{\frac{1}{2}}(U)(1-\delta^2)^{\frac{1}{2}}T_1), \qquad (۱)$$

باشد (باسو و همکاران، ۲۰۱۰) که در آن $T = k^{\frac{1}{2}}(U)|T_0|$، λ پارامتر چولگی و $\delta = \frac{\lambda}{(1+\lambda^2)^{\frac{1}{2}}}$. تابع چگالی Y به صورت

$$f(y) = 2\int_0^\infty \phi(y;\mu,u^{-1}\sigma^2)\Phi(\frac{u^{\frac{1}{2}}\lambda(y-\mu)}{\sigma})dH(u;\nu) \qquad (۲)$$

به دست می‌آید (گری و همکاران، ۲۰۱۱) که در آن $\phi(.;\mu,u^{-1}\sigma^2)$ تابع چگالی توزیع نرمال یک متغیره با میانگین μ و واریانس $u^{-1}\sigma^2 > 0$ و $\Phi(.)$ تابع توزیع نرمال استاندارد یک متغیره است. اگر $\lambda = 0$ آنگاه توزیع آمیخته– مقیاس نرمال حاصل می‌شود.

نتیجه ۲.۲. نمایش سلسله مراتبی متغیر تصادفی Y با توزیع $SMSN$ را می‌توان به صورت

$$\begin{aligned} Y \mid u,t &\sim N(\mu+\Delta t, k(u)\Gamma) \\ T \mid u &\sim TN(0,k(u))I(0,\infty) \qquad (۳) \\ U &\sim H(u,\nu) \end{aligned}$$

نشان داد که در آن $\Gamma = (1-\delta^2)\sigma^2$، $\Delta = \sigma\delta$ و $TN(0,k(u))I(0,\infty)$ نشان‌دهنده توزیع نرمال بریده شده با پارامترهای مکان صفر و مقیاس $k(u)$ است.

لم ۳.۲. فرض کنید $Y \sim SMSN(\mu,\sigma^2,\lambda;H)$

الف) اگر $E[U^{-\frac{1}{2}}] < \infty$، آنگاه $E[Y] = \mu + \sqrt{\frac{2}{\pi}}k_1\Delta$.

ب) اگر $E[U^{-1}] < \infty$، آنگاه $Var[Y] = \sigma^2k_2 - \frac{2}{\pi}k_1^2\Delta^2$ که در آن $k_m = E\{U^{-\frac{m}{2}}\}$ (باسو و همکاران، ۲۰۱۰).

اثبات. با استفاده از تعریف ۱.۲ و قاعده امید ریاضی مکرر اثبات به سادگی انجام می‌شود. □

با استفاده از لم ۳.۲ و در نظر گرفتن $\lambda = 0$ امید ریاضی و واریانس توزیع آمیخته–مقیاس نرمال به صورت $E[Y] = \mu$ و $Var[Y] = \sigma^2k_2$ هستند.

قضیه ٤.۲. اگر $Y \sim SMSN(\mu,\sigma^2,\lambda;H)$ آنگاه تابع مشخصه آن به صورت

$$\psi_Y(t) = e^{i\mu t}\int_0^\infty 2exp(-\tfrac{k(u)t^2\sigma^2}{2})\Phi(k(u)^{\frac{1}{2}}\Delta it)dH(u;\nu),$$

است (کیم و جنتن، ۲۰۱۱).

اثبات. با استفاده از تعریف ۱.۲، تغییر متغیر $t_0 - it\Delta k(u)^{\frac{1}{2}} = z$ و در نظر داشتن تقارن توزیع نرمال استاندارد حول صفر اثبات انجام می‌شود. □

قضیه ۵.۲. : اگر $Y \sim SMSN(\mu,\sigma^2,\lambda;H)$ و $z = ay+b$ که a و b اعداد ثابت هستند، آنگاه

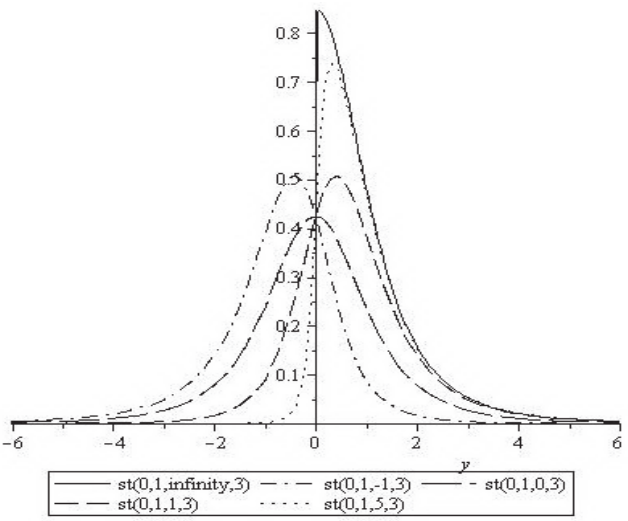
$$Z \sim SMSN(a\mu+b, a^2\sigma^2, \lambda; H)$$

اثبات. پس از محاسبه تابع مشخصه متغیر تصادفی Z با استفاده از قضیه ٤.۲ و در نظر داشتن خاصیت یکتایی تابع مشخصه، اثبات انجام می‌شود. با تغییر توزیع U در تعریف ۱.۲ انواع توزیع‌های $SMSN$ حاصل می‌شوند که در ادامه معرفی خواهند شد. □

توزیع چوله–تی: با قرار دادن $k(u) = \frac{1}{u}$ و فرض $U \sim Gamma(\frac{\nu}{2},\frac{\nu}{2})$ در عبارت (۱) توزیع چوله–تی (ST) با ν درجه آزادی حاصل می‌شود که با نماد $Y \sim ST(\mu,\sigma^2,\lambda;\nu)$ نشان داده می‌شود. شکل تابع چگالی آن به صورت

$$f(y) = 2\tfrac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}(1+\tfrac{d}{\nu})^{-\frac{\nu+1}{2}}T(\sqrt{\tfrac{\nu+1}{d+\nu}}A;\nu+1), \qquad y \in \mathbb{R}$$

است (گری و همکاران، ۲۰۱۱) که در آن $d = \frac{(y-\mu)^2}{\sigma^2}$، $A = \frac{\lambda(y-\mu)}{\sigma}$ و $T(.;\nu+1)$ مشخص کننده تابع توزیع تی–استیودنت با پارامترهای مکان صفر، مقیاس یک و $\nu+1$ درجه آزادی است. نمایش سلسله مراتبی نیز با جایگذاری توزیع U در نتیجه ۲.۲ برای این توزیع حاصل می‌شود. تابع چگالی این توزیع به ازای مقادیر مختلف λ در شکل ۱ نمایش داده شده است.



شکل ۱: تابع چگالی چوله-تی برای مقادیر مختلفλ

واضح است که اگر در این توزیع $\lambda = 0$ آنگاه توزیع تی-استیودنت با ν درجه آزادی حاصل می‌شود. در شکل۱ مشخص است که به ازای مقادیر منفی و مثبت برای λ چولگی به ترتیب به سمت چپ و راست است و به ازای مقادیر بزرگ پارامتر چولگی، شکل توزیع متمایل به تی بریده شده است. با تغییر مقدار درجه آزادی و در نظر گرفتن $\nu = 1$ و $\nu \to \infty$ به ترتیب توزیع‌های چوله-کوشی و چوله-نرمال حاصل می‌شود. با توجه به لم ۳.۲ اگر $Y \sim ST(\mu, \sigma^2, \lambda; \nu)$ آنگاه

الف) برای $\nu > 1$ $E[Y] = \mu + \sqrt{\frac{2}{\pi}} \frac{\Gamma(\frac{\nu-1}{2})(\frac{\nu}{2})^{\frac{1}{2}}}{\Gamma(\frac{\nu}{2})} \Delta,$.

ب) برای $\nu > 2$ $Var[Y] = \sigma^2 \frac{\nu}{\nu-2} - \frac{\nu}{\pi}(\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})})^2 \Delta^2,$.

قضیه ۶.۲. اگر $Y \sim ST(\mu, \sigma^2, \lambda; \nu)$ آنگاه تابع مشخصه آن به صورت

$$\psi_Y(t) = e^{i\mu t}\{\psi_T(t) + i\tau^*(\delta, \sigma t)\}$$

است که در آن

$$\tau^*(\delta, \sigma t) = \int_0^\infty e^{-\frac{t^2\sigma^2}{2u}} \tau(\Delta u^{-\frac{1}{2}} t) dH(u; \nu),$$

$$\tau^*(\delta, -\sigma t) = -\tau^*(\delta, \sigma t),$$

$$\tau(x) = \int_0^x \sqrt{\frac{2}{\pi}} e^{\frac{u^2}{2}} du \qquad x > 0, \quad \tau(-x) = -\tau(x)$$

و $\psi_T(t)$ تابع مشخصه توزیع تی-استیودنت به صورت

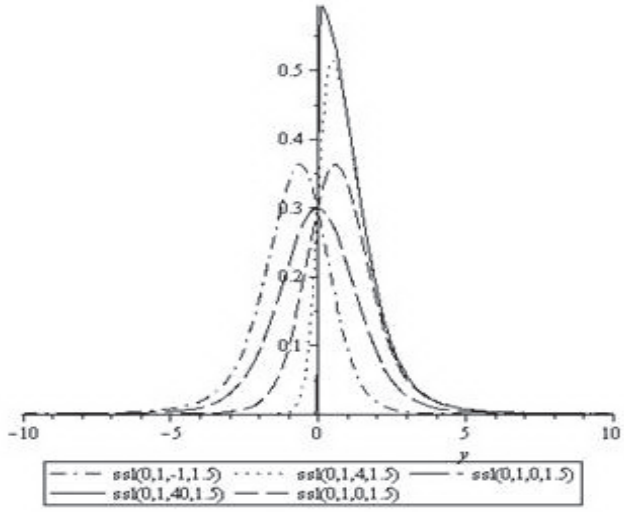
$$\psi_T(t) = \frac{B_{\frac{\nu}{2}}(\sigma\sqrt{\nu}|t|)(\sigma\sqrt{\nu}|t|)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2}) 2^{\frac{\nu}{2}-1}} \qquad t \in \mathbb{R}, \quad \nu > 0$$

است که $B_\alpha(b) = \frac{1}{2}\int_0^\infty x^{\alpha-1} e^{-\frac{1}{2}b(x+\frac{1}{x})} dx$ به ازای $a \in \mathbb{R}$ و $b > 0$ تابع بسل نوع دوم نامیده می‌شود(هارست، ۱۹۹۵).

توزیع چوله-اسلش: با فرض $k(u) = \frac{1}{u}$ و $U \sim Beta(\nu, 1)$ در عبارت (۱) توزیع چوله-اسلش (SSL) که یک توزیع دم-سنگین است با نماد $Y \sim SSL(\mu, \sigma^2, \lambda; \nu)$ به دست می‌آید(کانچو و همکاران، ۲۰۱۱). شکل تابع چگالی آن به صورت

$$f(y) = 2\nu \int_0^1 u^{\nu-1} \phi(y; \mu, u^{-1}\sigma^2) \Phi(u^{\frac{1}{2}} A) du, \qquad y \in \mathbb{R},$$

است.نمایش سلسله مراتبی نیز با جایگذاری توزیع U در نتیجه ۲.۲ برای این توزیع حاصل می‌شود. تابع چگالی به ازای مقادیر مختلف λ در شکل۲ آمده است.



شکل ۲: توزیع چوله-اسلش

با توجه به شکل۲ اگر در این توزیع $\lambda = 0$ آنگاه توزیع اسلش حاصل می‌شود و به ازای مقادیر بزرگ پارامتر چولگی، شکل توزیع متمایل به اسلش بریده شده است. با توجه به لم ۳.۲ اگر $Y \sim SSL(\mu, \sigma^2, \lambda; \nu)$ آنگاه

الف) برای $\nu > \frac{1}{2}$ $E[Y] = \mu + \sqrt{\frac{2}{\pi}} \frac{\nu}{\nu-\frac{1}{2}} \Delta,$

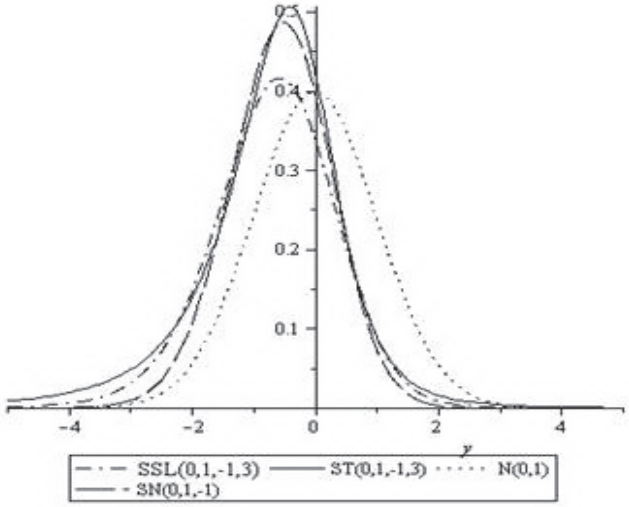
ب) برای $\nu > 1$ $Var[Y] = \sigma^2 \frac{\nu}{\nu-1} - \frac{2}{\pi}(\frac{\nu}{\nu-\frac{1}{2}})^2 \Delta^2,$

قضیه ۷.۲. اگر $Y \sim SSL(\mu, \sigma^2, \lambda; \nu)$ آنگاه تابع مشخصه آن به صورت

$$\psi_Y(t) = e^{i\mu t}\{m_T(\sigma t) + i\tau^*(\delta, \sigma t)\},$$

است که در آن

$$m_T(t) = \nu \int_0^1 e^{-\frac{t^2\sigma^2}{2u}} u^{\nu-1} du \qquad t \in \mathbb{R}, \quad \nu > 0$$



شکل ۳: مقایسه توزیع‌های متعلق به خانواده توزیع $SMSN$

در شکل۳ چهار توزیع متعلق به خانواده توزیع $SMSN$ به ازای پارامترهای مرکزی، مقیاس و چولگی یکسان رسم شده‌اند. در بین آن‌ها توزیع ST دم-سنگین‌تر است.

۳ مدل رگرسیونی غیرخطی

در مدل رگرسیونی غیرخطی پارامترهای مدل، غیرخطی هستند. در الگوی خطی، مشتق نسبت به یک پارامتر، مستقل از سایر پارامترها است در صورتی که در الگوهای غیرخطی، حداقل یکی از مشتقات، به یکی از پارامترها بستگی دارد.

تعریف ۱.۳. شکل کلی مدل رگرسیونی غیرخطی به صورت

$$y_i = \eta(\boldsymbol{\beta}, \boldsymbol{x}_i, \varepsilon_i), \qquad i = 1, ..., n \qquad (٤)$$

است و شکل متعارف‌تر آن با جمله خطای جمعی به صورت

$$y_i = \eta(\boldsymbol{\beta}, \boldsymbol{x}_i) + \varepsilon_i, \qquad i = 1, ..., n \qquad (٥)$$

و یا به شکل برداری $\boldsymbol{y} = \boldsymbol{\eta}(\boldsymbol{\beta}, \boldsymbol{X}) + \boldsymbol{\varepsilon}$ نوشته می‌شود که $\boldsymbol{\varepsilon}$ خطای تصادفی با میانگین صفر و واریانس $\sigma^2\boldsymbol{I}$ است (فکس، ۲۰۰۲). فرض معمول این است که توزیع $\boldsymbol{\varepsilon}$ نرمال باشد. یکی از مرحله‌های مهم در تحلیل غیرخطی مشخص کردن تابع انتظار $\boldsymbol{\eta}(\boldsymbol{\beta}, \boldsymbol{X})$ است که تابعی غیرخطی از $\boldsymbol{\beta}$ و یک به یک و دوبار مشتق پذیر نسبت به بردار ضرایب رگرسیونی $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$ است. تفاوت دو شکل نوشتار (٤) و (٥) در فرض‌های مربوط به جمله خطا است؛ مدل (٤) مدل‌هایی با خطای ضربی را نیز شامل می‌شود. مدل‌هایی را که با انجام تبدیل‌های مناسبی خطی می‌شوند مدل‌های "به طور طبیعی خطی" می‌نامند (باتز و واتز، ۱۹۸۸). باید توجه داشت که برآورد ضرایب رگرسیونی در مدل تبدیل یافته با مدل اولیه یکسان نیست، زیرا متغیرهای پاسخ و مستقل تبدیل شده‌اند. اهمیت مدل‌های خطی شدنی در رگرسیون غیرخطی استفاده از روش‌های برآوردیابی متداول و استنباط آماری معمول است. مدل‌هایی را که تحت هیچ تبدیلی خطی نشوند مدل‌های "به طور طبیعی غیرخطی" می‌نامند. در این مقاله مدل (٥) در نظر گرفته می‌شود. در این مدل، مجموع مربع‌های خطا به صورت

$$S(\boldsymbol{\beta}) = \Sigma_{i=1}^{n}(y_i - \eta(\boldsymbol{\beta}, \boldsymbol{x}_i))^2, \qquad i = 1, ..., n \qquad (٦)$$

تعریف می‌شود. برای برآورد پارامترها نیاز به معادله‌های نرمال داریم. از آن‌جا که معادله‌های مربوط به هر پارامتر به سایر پارامترها وابسته است، برای برآورد پارامترها به روش‌های تکراری نیاز است و در این روش مقدارهای آغازین مناسب وارد الگوریتم می‌شود و برآوردیابی ادامه می‌یابد تا در نهایت همگرایی رخ دهد. در بسیاری از موارد که مقدار اولیه نامناسب انتخاب شود ممکن است همگرایی ایجاد نشود. برخلاف رگرسیون خطی در این‌جا استنباط آماری توسط روش‌های مجانبی انجام می‌گیرد. با فرض نرمال و مستقل بودن توزیع خطاها، به کمک روش ماکسیمم درستنمایی نیز عبارت (٦) باید مینیمم شود. پس از محاسبه برآورد پارامترها ($\hat{\boldsymbol{\beta}}$) از طریق روش‌های مختلف برآوردیابی مانند تکرار گاوس-نیوتن، الگوریتم EM و الگوریتم نمونه‌گیر گیبز برآورد واریانس نیز از رابطه

$$s^2 = \frac{S(\hat{\boldsymbol{\beta}})}{n-p} = \frac{\sum_{i=1}^{n}(y_i - \eta(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i))^2}{n-p},$$

قابل محاسبه است. ماتریس کواریانس نیز به صورت

$$\widehat{Var(\hat{\boldsymbol{\beta}})} = s^2(W'W)^{-1},$$

قابل برآورد است که در آن W یک ماتریس $p \times p$ از مشتق اول $\eta(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i))$ نسبت به $\hat{\boldsymbol{\beta}}$ است که عناصر قطری آن واریانس هر پارامتر و سایر عناصر کواریانس بین برآورد پارامترها را نشان می‌دهند.

٤ مدل رگرسیون غیرخطی با توزیع خطای آمیخته-مقیاس چوله-نرمال

تعریف ۱.٤. اگر در معادله (۵) خطای تصادفی دارای توزیع $SMSN(-\sqrt{\frac{2}{\pi}}k_1\Delta, \sigma^2, \lambda; H)$ باشد، آنگاه مدل رگرسیون غیرخطی $SMSN$ تعریف می‌شود (گری و همکاران، ۲۰۱۱). با استفاده از قضیه ۵.۲ توزیع متغیر تصادفی Y_i برای $i = 1, ..., n$، به صورت

$$Y_i \sim SMSN(\eta(\boldsymbol{\beta}, \boldsymbol{x}_i) - \sqrt{\tfrac{2}{\pi}}k_1\Delta, \sigma^2, \lambda; H),$$

خواهد بود. همچنین توسط لم ۳.۲ امید ریاضی و واریانس متغیر پاسخ به صورت

$$E[Y_i] = \eta(\boldsymbol{\beta}, \boldsymbol{x}_i),$$
$$Var[Y_i] = \sigma^2 k_2 - b^2\Delta^2,$$

محاسبه می‌شوند که در آن $b = -\sqrt{\frac{2}{\pi}}k_1$. در این مدل اگر خطای تصادفی دارای توزیع آمیخته-مقیاس نرمال باشد آنگاه $Y_i \sim SMN(\eta(\boldsymbol{\beta}, \boldsymbol{x}_i), \sigma^2; H)$ و بنابراین

$$E[Y_i] = \eta(\boldsymbol{\beta}, \boldsymbol{x}_i),$$
$$Var[Y_i] = \sigma^2 k_2.$$

برآورد پارامترهای این مدل با استفاده از روش ماکسیمم درستنمایی نیاز به استفاده از روش‌های تکراری پیشرفته و یا روش‌های بیزی مبتنی بر $McMC$ دارد. ما در این مقاله از الگوریتم نمونه‌گیر گیبز (گلفند، ۲۰۰۰) که مبتنی بر $McMC$ است استفاده کرده‌ایم.

برای به کارگیری الگوریتم نمونه‌گیر گیبز ابتدا مدل رگرسیون غیرخطی $SMSN$ را به صورت سلسله مراتبی

$$\begin{aligned} Y_i|T_i = t_i &\sim N_1(\eta(\boldsymbol{\beta}, \boldsymbol{x}_i) + \Delta t_i, U_i^{-1}\Gamma) \\ T_i|U_i = u_i &\sim TN_1(b, u_i^{-1})I(b, \infty) \\ U_i &\sim H(., \boldsymbol{\nu}) \end{aligned} \quad (۷)$$

نمایش می‌دهیم. با فرض $\boldsymbol{y} = (y_1, ..., y_n)'$، $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_q)'$، $\boldsymbol{t} = (t_1, ..., t_n)'$ و $\boldsymbol{u} = (u_1, ..., u_n)'$ تابع درستنمایی کامل $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2, \lambda, \boldsymbol{\nu}')'$ به شرط $\boldsymbol{u}$، $\boldsymbol{t}$، $\boldsymbol{x}$ و $\boldsymbol{y}$ به صورت

$$L_c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{t}, \boldsymbol{u}) \propto \prod_{i=1}^{n} [\phi_1(y_i; \eta(\boldsymbol{\beta}, \boldsymbol{x}_i) + \Delta t_i, u_i^{-1}\Gamma)\phi_1(t_i; b, u_i^{-1})I(b, \infty)h(u_i|\boldsymbol{\nu})] \quad (۸)$$

نوشته می‌شود.

۱.٤ تحلیل بیزی مدل‌های رگرسیون غیرخطی با توزیع خطای $SMSN$

فرض کنید که پارامترهای مدل از یکدیگر مستقل باشند. به منظور سادگی انجام محاسبات بیزی، توزیع‌های پیشین آگاهی بخش ضعیف را برای پارامترهای مجهول $\boldsymbol{\beta}, \lambda, \sigma^2, \boldsymbol{\nu}$ به کار می‌بریم. توزیع پیشین برای ضرایب رگرسیونی نرمال با تابع چگالی $\pi(\beta_j) = \phi_1(\beta_j; \mu_{\beta_j}, \sigma^2_{\beta_j})$ به ازای $j = 1, ..., p$ و برای پارامترهای Γ و Δ به صورت

$$\Delta \sim N(\mu_\Delta, \sigma^2_\Delta), \qquad \Gamma^{-1}\ Gamma(\frac{\rho}{2}, \frac{\varrho}{2}) \quad (۹)$$

انتخاب شده‌اند که در آن ابر پارامترها معلوم در نظر گرفته شده‌اند. توزیع پیشین برای $\boldsymbol{\nu}$ با نماد $\pi(\boldsymbol{\nu})$ نشان داده می‌شود که شکل توزیع آن به توزیع $SMSN$ ویژه‌ای که مورد نظر است بستگی دارد. با این فرض‌هاچگالی پیشین توأم پارامترها به صورت

$$\pi(\boldsymbol{\theta}) = (\prod_{j=1}^{p} \pi(\beta_j; \mu_{\beta_j}, \sigma^2_{\beta_j}))\pi(\Delta; \mu_\Delta, \sigma^2_\Delta)\pi(\Gamma; \rho, \varrho)\pi(\boldsymbol{\nu}) \quad (۱۰)$$

نوشته می‌شود. چگالی پسین توأم با ضرب کردن روابط(۸) و(۹) و(۱۰) به صورت

$$\pi(\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{u}|\boldsymbol{y}, \boldsymbol{x}) \propto \prod_{i=1}^{n} [\phi_1(y_i; \eta(\boldsymbol{\beta}, \boldsymbol{x}_i) + \Delta t_i, u_i^{-1}\Gamma)\phi_1(t_i; b, u_i^{-1})I(b, \infty)h(u_i|\boldsymbol{\nu})]\pi(\boldsymbol{\theta}) \quad (۱۱)$$

به دست می‌آید. برای به دست آوردن چگالی پسین حاشیه‌ای هر پارامتر باید از رابطه (۱۱) نسبت به سایر پارامترها انتگرال گرفت، اما به دلیل پیچیده بودن محاسبات، از روش‌های $McMC$ مانند رهیافت نمونه‌گیر گیبز که تنها به توزیع‌های شرطی کامل نیاز دارد استفاده می‌شود. با انجام محاسبات جبری پیچیده و با ساده کردن عبارت (۱۱) نسبت به هر یک از پارامترها این توزیع‌ها به صورت زیر حاصل می‌شوند

$$T_i|\boldsymbol{\beta}, \Delta, \Gamma, \boldsymbol{\nu}, \boldsymbol{y}, \boldsymbol{u} \sim TN_1(\mu_{T_i} + b, u_i^{-1}M_T^2)I(b, \infty), \quad (۱۲)$$

$$\Delta|\boldsymbol{\beta},\boldsymbol{t},\Gamma,\boldsymbol{\nu},\boldsymbol{y},\boldsymbol{u}\sim N_1(\frac{\mathcal{D}\sigma_\Delta^2+\Gamma\mu_\Delta}{\zeta\sigma_\Delta^2+\Gamma},\frac{\zeta\sigma_\Delta^2+\Gamma}{\Gamma\sigma_\Delta^2}), \quad (۱۳)$$

$$\Gamma^{-1}|\boldsymbol{\beta},\boldsymbol{t},\Delta,\boldsymbol{\nu},\boldsymbol{y},\boldsymbol{u}\sim Gamma(\frac{1}{2}(m+n),\ \frac{1}{2}(nu+\sum_{i=1}^{n}u_i(y_i-\eta(\boldsymbol{\beta},\boldsymbol{x}_i)-\Delta t_i)^2)) \quad (۱۴)$$

$$\boldsymbol{\beta}|\Gamma,\boldsymbol{t},\Delta,\boldsymbol{\nu},\boldsymbol{y},\boldsymbol{u}\propto \phi_p(\boldsymbol{\beta};\boldsymbol{\mu_\beta},\boldsymbol{D})\prod_{i=1}^{n}\phi_1(\eta(\boldsymbol{\beta},\boldsymbol{x}_i);(y_i-\Delta t_i),u_i^{-1}\Gamma) \quad (۱۵)$$

$$\pi(u_i|\boldsymbol{\beta},\Gamma,\boldsymbol{t},\Delta,\boldsymbol{\nu},\boldsymbol{y},\boldsymbol{x})\propto u_i exp\{-\frac{1}{2\Gamma}u_i (y_i-\eta(\boldsymbol{\beta},\boldsymbol{x}_i)-\Delta t_i)^2-\frac{1}{2}u_i(t_i-b)^2\}h(u_i|\boldsymbol{\nu}) \quad (۱۶)$$

$$\pi(\boldsymbol{\nu}|\boldsymbol{\beta},\Gamma,\boldsymbol{t},\Delta,\boldsymbol{u},\boldsymbol{y})\propto\pi(\boldsymbol{\nu})\prod_{i=1}^{n}h(u_i|\boldsymbol{\nu}) \quad (۱۷)$$

که در این توزیع‌ها روابط

$\mu_{T_i}=\frac{\Delta}{\Gamma+\Delta^2}(y_i-\eta(\boldsymbol{\beta},\boldsymbol{x}_i)-b\Delta)$ و $M_T^2=\frac{\Delta}{\Gamma+\Delta^2}$،

$\mathcal{D}=\sum_{i=1}^{n}u_i t_i(y_i-\eta(\boldsymbol{\beta},\boldsymbol{x}_i))$ و $\zeta=\sum_{i=1}^{n}u_i t_i^2$،

$\boldsymbol{D}=diag(\sigma_{\beta_1}^2,...,\sigma_{\beta_p}^2)$ و $\boldsymbol{\mu_\beta}=(\mu_{\beta_1},...,\mu_{\beta_p})'$،

برقرار هستند. همانطور که مشخص است برای عبارت (۱۵) توزیع شناخته شده‌ای حاصل نشده است. بنابراین برای تولید نمونه به الگوریتم متروپلیس-هستینگز نیاز است. توزیع‌های پسین (۱۶) و (۱۷) به توزیع $SMSN$ ویژه‌ای که استفاده می‌شود و به چگالی پیشین $\boldsymbol{\nu}$ بستگی دارند.

به عنوان مثال مدل‌های زیر را در نظر بگیرید.

مدل چوله-تی: برای مدل رگرسیون غیرخطی با توزیع خطای ST، توزیع پیشین درجه آزادی، نمایی بریده شده در فاصله $(2,\infty)$ با میانگین اولیه $\frac{2}{\psi}$ در نظر گرفته می‌شود که با نماد $\nu\sim exp(\frac{\psi}{2})I(2,\infty)$ نشان داده می‌شود (کانچو و همکاران، ۲۰۱۱). بنابراین توزیع‌های شرطی کامل (۱۶) و (۱۷) به صورت

$$U_i|\boldsymbol{\beta},\Gamma,\boldsymbol{t},\Delta,\nu,\boldsymbol{y}\sim Gamma(\frac{\nu}{2}+1,\frac{\nu}{2}+\frac{\mathcal{C}_i}{2}),$$

$$\pi(\nu|\boldsymbol{\beta},\Gamma,\boldsymbol{t},\Delta,\boldsymbol{u},\boldsymbol{y})\propto \pi_1(\nu)\times Gamma(\frac{n\nu}{2}-1,\frac{1}{2}\sum_{i=1}^{n}(u_i-logu_i))I(2,\infty),$$

محاسبه می‌شوند که در آن $\mathcal{C}_i=\frac{1}{\Gamma}(y_i-\eta(\boldsymbol{\beta},\boldsymbol{x}_i)-\Delta t_i)^2+(t_i-b)^2$ و $\pi_1(\nu)=(2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2}))^{-n}$. توزیع شرطی کامل ν شکل مشخصی ندارد بنابراین برای تولید نمونه می‌توان از الگوریتم متروپلیس-هستینگز استفاده کرد.

مدل چوله-اسلش: برای مدل رگرسیون غیرخطی با توزیع خطای SSL، توزیع پیشین پارامتر ν را به صورت $Gamma(a,b)$ در نظر می‌گیریم که در آن a و b مقادیر مثبت و کوچک هستند و $b<<a$ (کانچو و همکاران، ۲۰۱۱). بنابراین توزیع‌های شرطی کامل (۱۶) و (۱۷) به صورت

$U_i|\boldsymbol{\beta},\Gamma,\boldsymbol{t},\Delta,\nu,\boldsymbol{y}\sim Gamma(\nu+1,\frac{\mathcal{C}_i}{2})I(0,1)$,

$\nu|\boldsymbol{\beta},\Gamma,\boldsymbol{t},\Delta,\boldsymbol{y},\boldsymbol{u}\sim Gamma(n+a,b-\sum_{i=1}^{n}logu_i)$,

ساده می‌شوند.

۵ مثال کاربردی: تحلیل داده‌های فارماکوکینتیکز

علم فارماکوکینتیک (Pharmacokinetics) به مطالعه عملکرد دارو در بدن می‌پردازد. داده‌های مورد استفاده (پین‌هیرو و بیتز، ۱۹۹۵) متعلق به مطالعات فارماکوکینتیک هستند و به دو گروه تقسیم شده‌اند، به گروه اول میزان دُز ٤ و به گروه دوم میزان دُز ۵ از داروی خوراکی تئوفیلین داده شده است. برای این داده‌ها غلظت دارو در مدت ۲۵ ساعت به عنوان متغیر پاسخ اندازه‌گیری می‌شود. مدل پیشنهاد شده برای ۹۸ نفر نمونه به صورت

$$y_i=\frac{dose_i ke_i ka_i}{cl_i/(k_{ai}-k_{ei})}(exp(-ke_i time_i)-exp(-ka_i time_i))+\varepsilon_i,\qquad i=1,..,98$$

ارائه شده است که در آن dose میزان دُز دارو و دارای دو مقدار ٤ و ۵ است که گروه اول دارو با دُز ٤ و گروه دوم با دُز ۵ را مصرف کرده‌اند. time زمان اندازه‌گیری متغیر پاسخ از صفر تا ۲۵ ساعت است. ke نرخ از بین رفتن در هر گروه است. ka میزان جذب دارو در هر گروه است. cl میزان حجم توزیع دارو در هر گروه است. همچنین مقادیر ke، ka و cl باید مثبت باشند. بدین منظور تبدیل‌های

$$k_{ei}=exp(\beta_3 group_i)$$

$$k_{ai}=exp(\beta_2 group_i)\quad \text{و}$$

$$cl_i=exp(\beta_1{}^* group_i)\quad \text{و}$$

را برای $i=1,..,98$ انجام می‌دهیم که در آن $\beta_1{}^*=\beta_1+\beta_{1diff}$. در مطالعات قبلی توزیع نرمال را در برازش مدل فوق برای

خطاها فرض کرده‌اند. ما ابتدا مدل فوق را با فرض خطای نرمال برازش داده و مانده‌ها را به دست آوردیم. از آنجا که آزمون شاپیرو ویلک نشان داد که توزیع مانده‌ها نرمال نیست چند توزیع از خانواده SMSN، از جمله چوله-تی، چوله-نرمال و چوله-اسلش را برای توزیع مانده‌ها در نظر گرفته‌ایم. علاوه بر آن توزیع‌های تی، اسلش و نرمال نیز برای مقایسه مدل‌ها در نظر گرفته شدند. برای انتخاب بهترین مدل برازش شده از معیارهای اطلاع بیزی (BIC) و معیار اطلاع آکائیک (AIC) به صورت $BIC = D(\boldsymbol{\theta})+2p_d ln(N)$ و $AIC = D(\boldsymbol{\theta})+2p_d$ استفاده کردیم (زفراس، ۲۰۰۹) که در آن p_d معرف تعداد پارامترهای مدل و N تعداد کل مشاهده‌ها و $D(\boldsymbol{\theta}) = -2loglikelihood(\boldsymbol{\theta}|data)$ معیار انحراف است. مدل با AIC و BIC کمتر به عنوان مدل بهتر انتخاب می‌شود. با توجه به اینکه از پیش در مورد توزیع ضرایب رگرسیونی اطلاعی در دست نیست برای آن‌ها توزیع پیشین $U(-100, 100)$ در نظر گرفته شده است. سایر پیشین‌ها به صورت $\Gamma^{-1} \sim Gamma(0.001, 0.001)$ ،$\nu \sim exp(0.1)I(2, \infty)$ در توزیع‌های تی و ST، $\nu \sim Gamma(0.1, 0.1)I(1, \infty)$ در توزیع‌های اسلش و SSL و $\Delta \sim N(0, 0.01)$ برای توزیع‌های چوله تعریف شده‌اند. نتایج برازش مدل‌ها و برآورد بیز پارامترهای مدل پس از همگرایی، با ۵۰۰۰۰ نمونه شبیه‌سازی شده و دورریز ۲۰۰۰۰ مقدار اولیه با استفاده از الگوریتم نمونه‌گیرگیبز توسط نرم‌افزار اُپن باگز (اسپیگل هالتر و همکاران،۲۰۱۰) در جدول ۱ نشان داده شده است.

جدول ۱: نتایج برآورد پارامترهای مدل

پارامترها	نرمال	چوله-نرمال	تی	چوله-تی	اسلش	چوله-اسلش
β_1	-۳/۵۲۴ (۰/۰۶۸۸۲)	-۳/۵۶۸ (۰/۰۷۲۳۲)	-۳/۵۱۱ (۰/۰۶۵۶۴)	-۳/۵۵۹ (۰/۰۷۱۳)	-۳/۵۱۹ (۰/۰۶۷۵۲)	-۳/۵۷۴ (۰/۷۰۴۱)
β_{1diff}	۰/۴۳۶۴ (۰/۰۴۶۱۸)	۰/۳۸۳۵ (۰/۰۴۶۰۶)	۰/۴۴۲۷ (۰/۰۴۵۶۲)	۰/۳۹۸۵ (۰/۰۴۴۳۵)	۰/۴۴۱۳ (۰/۰۴۵۱۲)	۰/۳۸۲۵ (۰/۰۴۸۸۸)
β_2	۰/۴۲۴۴ (۰/۱۱۱)	۰/۴۰۱۱ (۰/۰۹۱۷۸)	۰/۳۶۹۶ (۰/۱۱۵۴)	۰/۳۶۱ (۰/۱۰۰۷)	۰/۳۸۸۲ (۰/۱۱۷۶)	۰/۴۱۸ (۰/۰۹۴۱۳)
β_3	-۲/۵۶ (۰/۰۹۸۶۶)	-۲/۶۵۲ (۰/۱۰۶۹)	-۲/۵۴ (۰/۰۹۴۴۸)	-۲/۶۲۲ (۰/۱۰۲۳)	-۲/۵۵۲ (۰/۰۹۷۳۱)	-۲/۶۶ (۰/۱۰۴۴)
λ	-	۳/۸۹۶ (۳/۳۸۳)	-	۲/۹۴۵ (۱/۶۲۸)	-	۳/۷۹۷ (۳/۲۸۲)
ν	-	-	۱۳/۸۳ (۹/۳۳۱)	۱۵/۲۲ (۷/۲۱۷)	۵/۵۶۶ (۵/۶۲۶)	۶/۵۷ (۴/۹۳۴)
σ^2	۱/۵۲۸ (۰/۲۲۷۵)	۳/۷۵۶ (۰/۹۱۶۴)	۱/۲۵ (۰/۲۴۳۵)	۲/۸۵۸ (۰/۸۵۴۴)	۱/۰۷۷ (۰/۳۰۱۹)	۲/۸۴۵ (۰/۹۵۸۵)
Var	۱/۵۲۸ (۰/۲۲۷۵)	۱/۵۹۵ (۰/۲۷۲۴)	۱/۶۱۳ (۰/۶۴۳۱)	۱/۶۹۷ (۰/۶۴۲۴)	۲/۷۳۴ (۱۸/۰۲)	۱/۸۹۳ (۲/۳۵)
AIC	۳۲۳/۳	۳۱۶/۳	۳۲۴/۷	۳۴۷/۴	۳۲۴/۷	۳۹۵/۶
BIC	۳۳۶/۲	۳۳۱/۸	۳۴۰/۲	۳۶۵/۵	۳۴۰/۲	۴۱۳/۷

اعداد داخل پرانتز انحراف استاندارد بیزی را نشان می‌دهند.

با توجه به نتایج جدول ۱ ملاحظه می‌شود که معیارهای محاسبه شده برای مدل با خطای چوله-نرمال کمتر از سایر مدل‌ها است، بنابراین می‌توان نتیجه گرفت که برای داده‌های تحلیل شده، در نظر گرفتن توزیع خطای چوله-نرمال مناسب‌تر است. با توجه به فواصل اعتبار، در هر شش مدل برآورد ضرایب رگرسیونی و همچنین ضریب چولگی در مدل‌های چوله در سطح ۰/ ۰۵ معنادار است.

٦ بحث و نتیجه‌گیری

در برازش مدل‌های رگرسیونی خطی و غیرخطی فرض متداول مبتنی بر نرمال بودن توزیع خطا است. در صورت برقرار نبودن این فرض، نتایج نادرستی از برآورد پارامترهای مدل حاصل می‌شود. بنابراین، جایگزینی توزیع‌های منعطف دیگر که برای تحلیل داده‌ها مناسب باشد ضروری است. در این مقاله خانواده توزیع $SMSN$ معرفی گردید و ویژگی‌های آماری آن مورد بررسی قرار گرفت. همچنین کاربرد آن در مدل‌های رگرسیونی غیرخطی ارائه گردید. با استفاده از توزیع‌هایی از این خانواده در تحلیل داده‌های فارماکوکینتیکز نتیجه گرفته شد که مدل جایگزین در تحلیل این داده‌ها مناسب‌تر است. در نهایت با کاربرد معیارهای انتخاب مدل مشخص شد که مدل غیرخطی با مانده‌های چوله-نرمال بیشتر برازنده این داده‌ها است.

۷ تشکر قدردانی

نویسندگان از سردبیر و هیئت داوران که با ارائه پیشنهادات سازنده موجب بهبود این مقاله شدند، سپاسگزاری می‌کنند.

مراجع

[۱] شکل آبادی، ر.، کاظمی، ا. (۱۳۹۱). توزیع‌های آمیخته-مقیاس نرمال و کاربرد آنها در برازش مدل رگرسیون پانلی. اندیشه آماری. **۳۳ (۱)**، ۱٤-۱.

[2] Basso, R.M., Lachos, V.H., Cabral, C.R.B. and Ghosh, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions, *Computational Statistics Data Analysis*, **54(12)**, 2926-2941.

[3] Bates, D. M. and Watts. D. G. (1988). *Nonlinear Regression Analysis and Its Applications*, New York: Wiley.

[4] Cancho, V. C., Dey, D. K., Lachos, V. H. and Andrade, M. G. (2011). Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: estimation and case influence diagnostics, *Computational Statistics and Data Analysis*, **55**, 588-602.

[5] Fox, J. (2002). Nonlinear regression and nonlinear least squares, Appendix to "*An R and S-PLUS Companion to Applied Regression*", Sage Publications, Ca.

[6] Garay, A. M., Lachos, V. H. Abanto-Valle, C.A. (2011). Nonlinear regression models based on scale mixtures of skew-normal distributions, *Journal of the Korean Statistical Society*, **40**, 115-124.

[7] Gelfand, A.E. (2000). Gibbs Sampling. *Journal of the American Statistical Association*, **95**, 1300-1304.

[8] Hurst, S. (1995). *The characteristic function of the Student t-distribution*, in: Financial Mathematics Research Report 006-95, Australian National University, Canberra ACT 0200, Australia.

[9] Kim, H.M. and Genton, M. G. (2011). Characteristic functions of scale mixtures of multivariate skew-normal distributions, *Journal of Multivariate Analysis*, **102(7)**, 1105-1117.

[10] Spiegelhalter, D. J. Thomas, A. Best, N. G. and Lunn, D.(2010). *OpenBugs User Manual, version 3.1.1*. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK, and Department of Mathematics and Statistics, University of St Andrews, Scotland, and Department of Epidemiology and Public Health, Imperial College School of Medicine, London, (www.mrc-bsu.cam.ac.uk/bugs).

[11] Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics*, **4**, 12–35.

[12] SAS Institute Inc. (2008) *SAS/STAT® 9.2 User's guide: the NLIN procedure*. Cary, NC: SAS Institute Inc.