

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 برآش مدل‌های رگرسیونی غیرخطی نرمال به منظور تحلیل داده‌ها با ساختار پخش متقارن به صورت توابع غیرخطی از پارامترهای مجھول 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-اسلس 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 برآش می‌دهیم تا اهمیت مدل پیشنهادی را بیان کنیم.

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

۱ مقدمه

در دهه اخیر تحقیق‌های متعددی جهت معرفی توزیع‌های آمیخته- مدل‌های رگرسیون غیرخطی در نظر گرفته و توسط روش‌های مقیاس نرمال به عنوان جایگزینی مناسب برای نرمال انجام شده بیزی پارامترهای مدل را برآورد می‌کنیم. ساختار این مقاله به است که از آن جمله می‌توان به شکل آبادی و کاظمی (۱۳۹۱) اشاره این صورت است که ابتدا در بخش دوم خانواده توزیع آمیخته-نمود. این توزیع‌ها در تحلیل مشاهده‌ها با ساختار پخش متقارن و مقیاس چوله-نرمال معروفی می‌شود. آنگاه در بخش سوم مدل دم سنگین به عنوان جایگزین نرمال به کار می‌رond. در این مقاله، رگرسیونی غیرخطی را همراه با برخی از ویژگی‌های خاص آن کاربرد خانواده جدیدی از توزیع‌های آمیخته-مقیاس چوله-نرمال معرفی می‌نماییم. در بخش چهارم مدل‌های رگرسیونی غیرخطی با کاربرد خانواده از توزیع‌های آمیخته-مقیاس چوله-نرمال (SMSN – NLM) را مورد بررسی (SMSN) را در برآش مدل‌های رگرسیونی غیرخطی بررسی می‌کنیم. برآش 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 گیز را به دست می‌آوریم. در بخش پنجم با استفاده از روش مارکفی (McMC) را در برآش مدل‌های فوق پیشنهاد داده‌اند. استفاده از رهیافت McMC و کاربردی از خانواده SMSN در با مانده‌های SMSN را روی داده‌های واقعی برآش داده و با 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 برآش

^۱دانشجوی کارشناسی ارشد دانشگاه اصفهان

^۲عضو هیأت علمی گروه آمار دانشگاه اصفهان

با استفاده از $\lambda = 3.2$ و در نظر گرفتن $E[Y] = \mu$ و $Var[Y] = \sigma^2 k_2$ واریانس توزیع آمیخته-مقیاس نرمال به صورت $\lambda = \frac{\mu}{\sigma^2 k_2}$ هستند.

قضیه ۴.۲. اگر $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$ آنگاه تابع مشخصه آن به صورت $\psi_Y(t) = e^{i\mu t} \int_0^\infty 2\exp(-\frac{k(u)t^2\sigma^2}{2})\Phi(k(u)^{\frac{1}{2}}\Delta it)dH(u; \nu)$, است (کیم و جتن، ۲۰۱۱).

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

قضیه ۵.۲. اگر $z = ay + b$ و $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$ 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 ۴.۲ و در نظر داشتن خاصیت یکتاپی تابع مشخصه، اثبات انجام می شود. با تغییر توزیع U در تعریف ۱.۲ انواع توزیع های $SMSN$ حاصل می شوند که در ادامه معرفی خواهند شد. \square

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 $Y \sim ST(\mu, \sigma^2, \lambda; \nu)$ درجه آزادی حاصل می شود که با نماد $N(\mu, \sigma^2, \lambda; \nu)$ نشان داده می شود. شکل تابع چگالی آن به صورت $f(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} (1 + \frac{d}{\nu})^{-\frac{\nu+1}{2}} T(\sqrt{\frac{\nu+1}{d+\nu}} A; \nu+1)$, $y \in \mathbb{R}$ است (گری و همکاران، ۲۰۱۱) که در آن $A = \frac{(y-\mu)^2}{\sigma^2}$, $d = \frac{\lambda(y-\mu)}{\sigma}$ و $T(\cdot; \nu+1)$ مشخص کننده تابع توزیع تی-استیومنت با پارامترهای مکان صفر، مقیاس یک و درجه آزادی $\nu+1$ در نتیجه است. نمایش سلسله مراتبی نیز با جایگذاری توزیع U در نتیجه ۲. برای این توزیع حاصل می شود. تابع چگالی این توزیع به ازای مقادیر مختلف λ در شکل ۱ نمایش داده شده است.

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 $H(\cdot; v)$ است که 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

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

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

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

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

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

$$Y | u, t \sim N(\mu + \Delta t, k(u)\Gamma)$$

$$T | u \sim TN(0, k(u))I(0, \infty) \quad (3)$$

$$U \sim H(u, \nu)$$

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

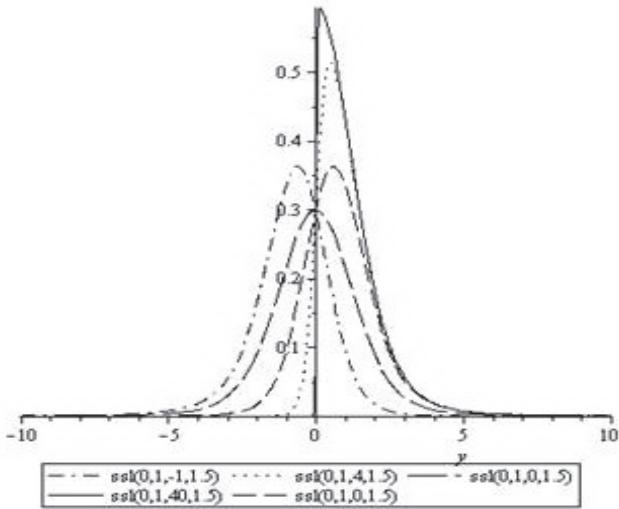
لم ۳.۲. فرض کنید $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$ الف) اگر $E[U^{-\frac{1}{2}}] < \infty$, آنگاه $E[U^{-\frac{1}{2}}] < \sqrt{\frac{2}{\pi}}k_1\Delta$, آنگاه $E[U^{-\frac{1}{2}}] < \sqrt{\frac{2}{\pi}}k_1\Delta$, آنگاه $E[U^{-1}] = \sigma^2 k_2 - \frac{2}{\pi}k_1^2\Delta^2$ که در آن $k_m = E\{U^{-\frac{m}{2}}\}$ (باسو و همکاران، ۲۰۱۰).

اثبات. با استفاده از تعریف ۱.۲ و قاعده اميد ریاضی مکرر اثبات به سادگی انجام می شود. \square

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 $(\mu, \sigma^2, \lambda; \nu)$ به $Y \sim SSL(\mu, \sigma^2, \lambda; \nu)$ دست می‌آید (کانچو و همکاران، ۲۰۱۱). شکل تابع چگالی آن به صورت

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

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 شکل ۲ آمده است.



شکل ۲: توزیع چوله-اسلش

با توجه به شکل ۲ 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

$$Y \sim SSL(\mu, \sigma^2, \lambda; \nu)$$

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

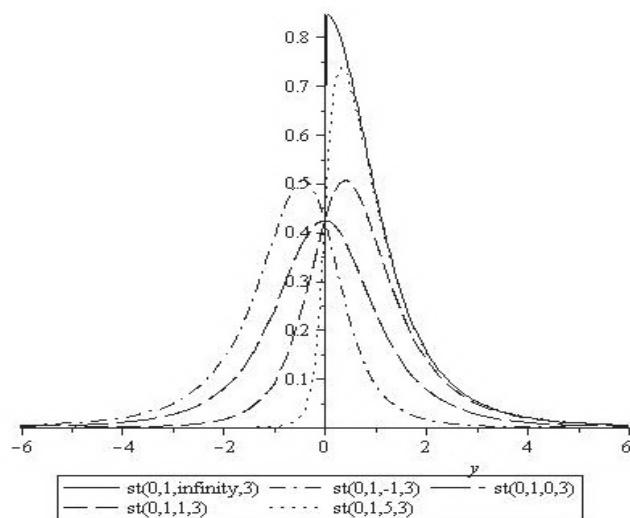
$$(b) \text{ برای } \nu > 1, \quad Var[Y] = \sigma^2 \frac{\nu}{\nu - 1} - \frac{2}{\pi} \left(\frac{\nu}{\nu - \frac{1}{2}} \right)^2 \Delta^2,$$

قضیه ۶.۲. 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

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



شکل ۱: تابع چگالی چوله-تی برای مقادیر مختلف λ

واضح است که اگر در این توزیع $\nu = 0$ آنگاه توزیع تی-استیوندنت با ν درجه آزادی حاصل می‌شود. در شکل ۱ 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 روشی چوله-نرمال حاصل می‌شود. با توجه به لم ۳.۲ اگر

$$Y \sim ST(\mu, \sigma^2, \lambda; \nu)$$

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

$$(b) \text{ برای } \nu > 2, \quad Var[Y] = \sigma^2 \frac{\nu}{\nu-2} - \frac{\nu}{\pi} \left(\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})} \right)^2 \Delta^2,$$

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

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

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

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

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

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

$$a \in \mathbb{R} \text{ به ازای } B_\alpha(b) = \frac{1}{2} \int_0^\infty x^{\alpha-1} e^{-\frac{1}{2}b(x+\frac{1}{x})} dx$$

و $b > 0$ تابع بدل نوع دوم نامیده می‌شود (هارست، ۱۹۹۵).

شامل می‌شود. مدل‌هایی را که با انجام تبدیل‌های مناسبی خطی می‌شوند مدل‌های "به طور طبیعی خطی" می‌نامند (باتز و واتر، 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 برآوردهایی متداول و استنباط آماری معمول است. مدل‌هایی را که تحت هیچ تبدیلی خطی نشوند مدل‌های "به طور طبیعی غیرخطی" می‌نامند. در این مقاله مدل (۵) در نظر گرفته می‌شود. در این مدل، مجموع مربع‌های خطأ به صورت

$$S(\beta) = \sum_{i=1}^n (y_i - \eta(\beta, x_i))^2, \quad i = 1, \dots, n \quad (۶)$$

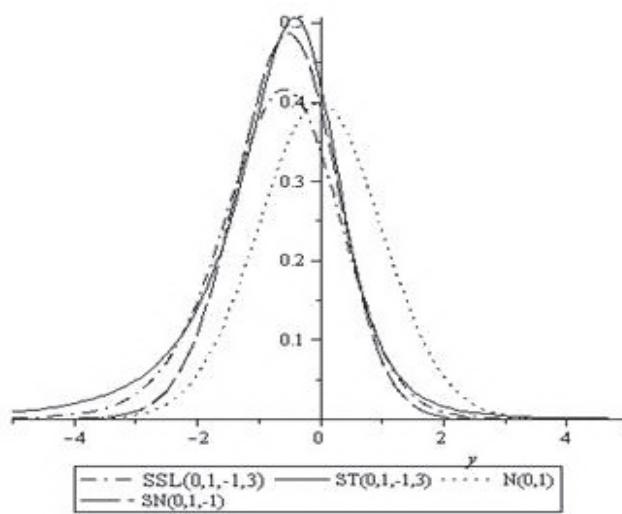
تعریف می‌شود. برای برآورد پارامترها نیاز به معادله‌های نرمال داریم. از آنجا که معادله‌های مربوط به هر پارامتر به سایر پارامترها وابسته است، برای برآورد پارامترها به روش‌های تکراری نیاز است و در این روش مقدارهای آغازین مناسب وارد الگوریتم می‌شود و برآوردهایی ادامه می‌یابد تا در نهایت همگرایی رخ دهد. در بسیاری از موارد که مقدار اولیه نامناسب انتخاب شود ممکن است همگرایی ایجاد نشود. برخلاف رگرسیون خطی در اینجا استنباط آماری توسط روش‌های مجذبی انجام می‌گیرد. با فرض نرمال و مستقل بودن توزیع خطأها، به کمک روش ماکسیمم درستنمایی نیز عبارت (۶) باید مینیمم شود. پس از محاسبه برآورد پارامترها ($\hat{\beta}$) از طریق روش‌های مختلف برآوردهایی مانند تکرار گالوس-نیوتون، الگوریتم EM و الگوریتم نمونه‌گیری گیز برآورد واریانس نیز از رابطه

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

قابل محاسبه است. ماتریس کواریانس نیز به صورت

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

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



شکل ۳: مقایسه توزیع‌های متعلق به خانواده توزیع SMSN در شکل ۳ چهار توزیع متعلق به خانواده توزیع SMSN به ازای پارامترهای مرکزی، مقیاس و چولگی یکسان رسم شده‌اند. در بین آن‌ها توزیع ST دم‌سنگین‌تر 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 مشتق‌ها، به یکی از پارامترها بستگی دارد.

تعریف ۱.۳. شکل کلی مدل رگرسیونی غیرخطی به صورت

$$y_i = \eta(\beta, x_i, \varepsilon_i), \quad i = 1, \dots, n \quad (۴)$$

است و شکل متعارف‌تر آن با جمله خطای جمعی به صورت

$$y_i = \eta(\beta, x_i) + \varepsilon_i, \quad i = 1, \dots, n \quad (۵)$$

و یا به شکل برداری $\varepsilon = \eta(\beta, X) + \varepsilon$ نوشته می‌شود که ε خطای تصادفی با میانگین صفر و واریانس $\sigma^2 I$ است (فکس، ۲۰۰۲). فرض معمول این است که توزیع ε نرمال باشد. یکی از مرحله‌های مهم در تحلیل غیرخطی مشخص کردنتابع انتظار $\eta(\beta, X)$ است که تابعی غیرخطی از β و یک به یک و دوبار مشتق پذیر نسبت به بردار ضرایب رگرسیونی $(\beta_0, \dots, \beta_k)' = \beta$ است. تفاوت دو شکل نوشتار (۴) و (۵) در فرض‌های مربوط به جمله خطأ است؛ مدل (۴) مدل‌هایی با خطای ضربی را نیز می‌دهند.

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

$$L_c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}, \mathbf{t}, \mathbf{u}) \propto \prod_{i=1}^n [\phi_1(y_i; \eta(\boldsymbol{\beta}, \mathbf{x}_i) + \Delta t_i, u_i^{-1}\Gamma)\phi_1(t_i; b, u_i^{-1})I(b, \infty)h(u_i|\boldsymbol{\nu})] \quad (8)$$

نوشته می‌شود.

۱.۴ 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 مجھول $\boldsymbol{\beta}, \lambda, \sigma^2, \nu$ به کار می‌بریم. توزیع پیشین برای ضرایب رگرسیونی نرمال با تابع چگالی $\pi(\boldsymbol{\beta}_j) = \phi_1(\beta_j; \mu_{\beta_j}, \sigma_{\beta_j}^2)$ به ازای $j = 1, \dots, p$ و برای پارامترهای Γ و Δ به صورت

$$\Delta \sim N(\mu_\Delta, \sigma_\Delta^2), \quad \Gamma^{-1} \sim \text{Gamma}\left(\frac{\rho}{2}, \frac{\varrho}{2}\right) \quad (9)$$

انتخاب شده‌اند که در آن ابر پارامترها معلوم در نظر گرفته شده‌اند. توزیع پیشین برای ν با نماد (ν) نشان داده می‌شود که شکل توزیع آن به توزیع SMSN ویژه‌ای که مورد نظر است بستگی دارد. با این فرض چگالی پیشین توأم پارامترها به صورت

$$\pi(\boldsymbol{\theta}) = \left(\prod_{j=1}^p \pi(\beta_j; \mu_{\beta_j}, \sigma_{\beta_j}^2) \right) \pi(\Delta; \mu_\Delta, \sigma_\Delta^2) \pi(\Gamma; \rho, \varrho) \pi(\nu) \quad (10)$$

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

$$\pi(\boldsymbol{\theta}, \mathbf{t}, \mathbf{u}|\mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^n [\phi_1(y_i; \eta(\boldsymbol{\beta}, \mathbf{x}_i) + \Delta t_i, u_i^{-1}\Gamma)\phi_1(t_i; b, u_i^{-1})I(b, \infty)h(u_i|\boldsymbol{\nu})]\pi(\boldsymbol{\theta}) \quad (11)$$

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 نمونه‌گیری گیز (گلفند، ۲۰۰۰) که مبتنی بر McMC است استفاده اسفلاده 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

$$T_i|\boldsymbol{\beta}, \Delta, \Gamma, \nu, \mathbf{y}, \mathbf{u} \sim TN_1(\mu_{T_i} + b, u_i^{-1}M_T^2)I(b, \infty), \quad (12)$$

تعريف ۱.۴. 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 $i = 1, \dots, n$ به صورت

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

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

$$E[Y_i] = \eta(\boldsymbol{\beta}, \mathbf{x}_i),$$

$$Var[Y_i] = \sigma^2 k_2 - b^2 \Delta^2,$$

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 $Y_i \sim SMN(\eta(\boldsymbol{\beta}, \mathbf{x}_i), \sigma^2; H)$

$$E[Y_i] = \eta(\boldsymbol{\beta}, \mathbf{x}_i),$$

$$Var[Y_i] = \sigma^2 k_2.$$

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

برای به کارگیری الگوریتم نمونه‌گیری گیز ابتدا مدل رگرسیون غیرخطی SMSN را به صورت سلسله مراتبی

$$Y_i|T_i = t_i \sim N_1(\eta(\boldsymbol{\beta}, \mathbf{x}_i) + \Delta t_i, U_i^{-1}\Gamma)$$

$$T_i|U_i = u_i \sim TN_1(b, u_i^{-1})I(b, \infty) \quad (V)$$

$$U_i \sim H(., \boldsymbol{\nu})$$

نمایش می‌دهیم. با فرض $\boldsymbol{\beta} = (x_1, \dots, x_q)'$, $\mathbf{y} = (y_1, \dots, y_n)'$ ، $\mathbf{u} = (u_1, \dots, u_n)'$ و $\mathbf{t} = (t_1, \dots, t_n)'$ تابع درستنمایی کامل

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 $a << b$ (کانچو و همکاران, ۲۰۱۱). بنابراین توزیع‌های شرطی کامل (۱۶) و (۱۷) به صورت

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

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

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 ۴ 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

$$y_i = \frac{dose \cdot ke \cdot ka_i}{cl_i / (ka_i - ke)} (exp(-ke \cdot time_i) - exp(-ka_i \cdot time_i)) + \varepsilon_i, \quad i = 1, \dots, 98$$

ارائه شده است که در آن dose میزان دُز دارو و دارای دو مقدار ۴ و ۵ است که گروه اول دارو با دُز ۴ 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

$$ke_i = exp(\beta_3 group_i)$$

$$ka_i = exp(\beta_2 group_i)$$

$$cl_i = exp(\beta_1 * group_i)$$

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

$$\Delta|\beta, t, \Gamma, \nu, y, u \sim N_1\left(\frac{\mathcal{D}\sigma_\Delta^2 + \Gamma\mu_\Delta}{\zeta\sigma_\Delta^2 + \Gamma}, \frac{\zeta\sigma_\Delta^2 + \Gamma}{\Gamma\sigma_\Delta^2}\right), \quad (13)$$

$$\Gamma^{-1}|\beta, t, \Delta, \nu, y, u \sim Gamma\left(\frac{1}{2}(m+n)\right),$$

$$\frac{1}{2}(nu + \sum_{i=1}^n u_i(y_i - \eta(\beta, x_i) - \Delta t_i)^2)) \quad (14)$$

$$\beta|\Gamma, t, \Delta, \nu, y, u \propto$$

$$\phi_p(\beta; \mu_\beta, D) \prod_{i=1}^n \phi_1(\eta(\beta, x_i); (y_i - \Delta t_i), u_i^{-1}\Gamma) \quad (15)$$

$$\pi(u_i|\beta, \Gamma, t, \Delta, \nu, y, x) \propto u_i \exp\left\{-\frac{1}{2\Gamma} u_i\right.$$

$$(y_i - \eta(\beta, x_i) - \Delta t_i)^2 - \frac{1}{2} u_i(t_i - b)^2\} h(u_i|\nu) \quad (16)$$

$$\pi(\nu|\beta, \Gamma, t, \Delta, u, y) \propto \pi(\nu) \prod_{i=1}^n h(u_i|\nu) \quad (17)$$

که در این توزیع‌ها روابط

$$\mu_{T_i} = \frac{\Delta}{\Gamma + \Delta^2}(y_i - \eta(\beta, x_i) - b\Delta) \quad \text{و} \quad M_T^2 = \frac{\Delta}{\Gamma + \Delta^2},$$

$$D = \sum_{i=1}^n u_i t_i (y_i - \eta(\beta, x_i)) \quad \text{و} \quad \zeta = \sum_{i=1}^n u_i t_i^2,$$

$$D = diag(\sigma_{\beta_1}^2, \dots, \sigma_{\beta_p}^2) \quad \text{و} \quad \mu_\beta = (\mu_{\beta_1}, \dots, \mu_{\beta_p})'$$

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 پیشنهاد شده است. پس از این توزیع‌ها روابط

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 $\exp(\frac{\psi}{2})I(2, \infty) \sim \nu$ نشان داده می‌شود (کانچو و همکاران, ۲۰۱۱). بنابراین توزیع‌های شرطی کامل (۱۶) و (۱۷) به صورت

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

$$\pi(\nu|\beta, \Gamma, t, \Delta, u, y) \propto$$

$$\pi_1(\nu) \times Gamma(\frac{n\nu}{2} - 1, \frac{1}{2} \sum_{i=1}^n (u_i - logu_i)) I(2, \infty),$$

محاسبه می‌شوند که در آن $C_i = \frac{1}{\Gamma}(y_i - \eta(\beta, x_i) - \Delta t_i)^2 + (t_i - b)^2$ 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

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 (۰/۰۷۰۴۱)
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)	-۲/۵۰۲ (۰/۰۹۷۳۱)	-۲/۶۶ (۰/۱۰۴۴)
λ	- (۳/۳۸۳)	۳/۸۹۶ (۳/۳۸۳)	- (۱/۶۲۸)	۲/۹۴۵ (۱/۶۲۸)	- (۵/۰۶۶)	۳/۷۹۷ (۳/۲۸۲)
ν	-	-	۱۳/۸۳ (۹/۳۳۱)	۱۵/۲۲ (۷/۲۱۷)	۵/۰۶۶ (۵/۷۱۶)	۶/۵۷ (۴/۹۳۴)
σ^2	۱/۰۲۸ (۰/۲۲۷۵)	۳/۷۵۶ (۰/۹۱۶۴)	۱/۲۵ (۰/۲۴۳۵)	۲/۸۵۸ (۰/۸۵۴۴)	۱/۰۷۷ (۰/۳۰۱۹)	۲/۸۴۵ (۰/۹۵۸۵)
Var	۱/۰۲۸ (۰/۲۲۷۵)	۱/۰۹۵ (۰/۲۷۷۲۴)	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 می دهند.

تحلیل داده‌ها مناسب باشد ضروری است. در این مقاله خانواده توزیع *SMSN* معرفی گردید و ویژگی‌های آماری آن مورد بررسی 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. تحلیل این داده‌ها مناسب‌تر است. در نهایت با کاربرد معیارهای انتخاب مدل مشخص شد که مدل غیرخطی با مانده‌های چوله-نرمال بیشتر برآزنده این داده‌ها است.

۶ بحث و نتیجه‌گیری

۷ تشکر قدردانی

در برآذش 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 نویسنده‌گان از سردبیر و هیئت داوران که با ارائه پیشنهادات سازنده می‌شود. بنابراین، جایگزینی توزیع‌های منعطف دیگر که برای موجب بهبود این مقاله شدند، سپاسگزاری می‌کنند.

مراجع

- [1] شکل آبادی، ر.، کاظمی، ا. (۱۳۹۱). توزیع‌های آمیخته-مقیاس نرمال و کاربرد آنها در برآذش مدل رگرسیون پانلی. *اندیشه آماری*. ۱۴(۱)، ۳۳-۱۴.
- [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.