توزیع چوله-اسلش و کاربرد آن در مباحث رگرسیونی

زهرانیکنام [۱] ، **محمدحسین علامت ساز** [۲]

چکیده:

فرض متداول در بسیاری از تحلیل داده‌های آماری نرمال بودن توزیع مشاهدات است. امّا این فرض غالباً در تحلیل داده‌های واقعی نقض می‌شود. بدین منظور توزیع‌های انعطاف پذیری به عنوان جایگزین توزیع نرمال پیشنهاد شده است. در این رابطه می‌توان به توزیع اسلش و چوله-اسلش اشاره کرد که در دهه‌ی اخیر از سوی محققان زیادی مورد توجه قرار گرفته اند. توزیع اسلش به عنوان یک توزیع دم-سنگین و متقارن در مطالعات استوار شناخته شده است. امّا با توجه به اینکه در مثالهای تجربی، مواقع زیادی وجود دارد که توزیع های متقارن برای برازش داده‌ها مناسب نمی‌باشد مطالعه‌ی توزیع ها در حالت چوله نیز از اهمیت ویژه‌ای برخوردار است، از این رو معرفی تعمیم‌های چوله‌ای از توزیع اسلش مورد توجه محققان قرار گرفته است. کاربردهایی از این توزیع به طور معمول در سیستم‌های چند مؤلفه‌ای پیچیده ای مانند سیستم‌های زیست محیطی، اقتصاد، جامعه شناسی، امور مالی و ... یافت می‌شود. در این مقاله ضمن بررسی توزیع چوله-اسلش و ویژگی‌های آن با استفاده از مجموعه داده‌های واقعی به بررسی اهمیت کاربرد این توزیع در مدل‌های رگرسیونی می‌پردازیم.

واژه‌های کلیدی: آمیخته مقیاس چوله- نرمال، برآورد کردن، توزیع دم سنگین، توزیع چوله-اسلش.

۱ مقدمه

در مدل‌های رگرسیون خطی معمولاً فرض بر این است که خطاها مستقل و دارای توزیع نرمال می‌باشند، امّا در اکثر مثال‌ها و داده‌های واقعی حالت‌هایی وجود دارد که این مدل به عنوان یک توزیع متقارن، ممکن است برای داده‌ها برازش مناسبی نباشد. توزیع اسلش به عنوان یک توزیع دم-سنگین و متقارن در مطالعات استوار شناخته شده است. در توصیف بسیاری از موارد که توزیع‌های متقارن قادر به توصیف آنها نیستند توزیع‌های چوله می‌توانند بسیار مفید واقع شوند. از این رو در تحقیقات برخی از توزیع‌های چوله به عنوان یک مدل انعطاف پذیر معرفی شده‌اند. یکی از توزیع‌هایی که می‌توانند جانشین مناسبی برای توزیع نرمال باشند توزیع چوله-اسلش است که این توزیع از توزیع نرمال دم-سنگین تر و دارای کشیدگی بیشتر و همچنین چولگی می‌باشد. در این مقاله به معرفی مدل‌های رگرسیونی خطی با خطای توزیع اسلش و چوله-اسلش می‌پردازیم. توزیع اسلش و بررسی ویژگی های آن توسط محققان زیادی از جمله راجرز و توکی [۳] [۱٤]، مورجنتالر [٤] [۱۳]، جمشیدیان [۱۰]، کاشید و کالکارنی [۵] [۱۲] صورت گرفته است. سپس کفادار [٦] [۱۱] برآورد حداکثر درستنمایی پارامترهای مقیاسی و مکانی برای توزیع اسلش استاندارد به دست آورد. وانگ و جنتون [۷] [۱٦] تعریف دیگری از توزیع چوله-اسلش چند متغیره را بر اساس تعریفی که آزالینی [۸] [۲] برای توزیع چوله-نرمال ارائه

[۱] کارشناس ارشد آمار، دانشگاه اصفهان

[۲] عضو هیئت علمی آمار، دانشگاه اصفهان

[۳] Rogers and Tukey

[٤] Morgenthaler

[۵] Kashid and Kulkarni

[٦] Kafadar

[۷] Wang and Genton

[۸] Azzalini

داد، معرفی کردند. آن‌ها همچنین توزیع چوله –اسلش چند متغیره را بر اساس آمیخته–مقیاسی از توزیع‌های نرمال و یکنواخت ارائه و ویژگی‌های آن را مورد بررسی قرار دادند. کریستینا و همکاران [۹] [٦] مدل سازی سری های زمانی مالی را با استفاده از توزیع چوله–اسلش انجام دادند. با توجه به اینکه توزیع چوله–اسلش عضو خانواده توزیع های آمیخته مقیاس چوله–نرمال است. در بخش ۲ خانواده توزیع های آمیخته مقیاس چوله–نرمال را معرفی می‌کنیم. در بخش ۳ توزیع اسلش را معرفی و ویژگی‌های آن را بررسی می‌کنیم. در بخش ٤ به معرفی توزیع چوله–اسلش و ویژگی های آن می‌پردازیم و در بخش ۵ مدل رگرسیون خطی را معرفی می‌کنیم و در بخش٦ مدل های رگرسیونی خطی را با خطای توزیع چوله–اسلش معرفی می‌کنیم و با استفاده از مجموعه داده‌های واقعی که تنها با استفاده از مدل رگرسیونی نرمال در متون مختلف مرتبط با مباحث رگرسیونی برازش داده شده است. ما با جایگزینی توزیع چوله–اسلش به بررسی اهمیت کاربرد توزیع چوله–اسلش در مدل‌های رگرسیونی می‌پردازیم. در پایان نتیجه گیری را ارائه می‌دهیم.

۲ خانواده توزیع آمیخته–مقیاس چوله–نرمال (SMSN)

خانواده توزیع آمیخته – مقیاس چوله – نرمال $(SMSN)$ [۱۰] که تعمیمی از توزیع چوله–نرمال است با لحاظ کردن چولگی و دم سنگین بودن در تحلیل داده‌ها، توزیع‌های بسیاری از جمله چوله–تی (ST) [۱۱] و چوله–اسلش (SSL) [۱۲] را شامل می‌شود. در تحلیل داده‌هایی که مدل‌های رگرسیون با فرض نرمال بودن مؤلفه‌های خطا به آن‌ها خوب برازش نمی شوند، اهمّیت استفاده از توزیع‌های دیگر خانواده توزیع آمیخته–مقیاس چوله–نرمال روشن است. اهمّیت ویژه خانواده معرفی شده در کاربرد آنها جهت مدل سازی داده‌هایی با ساختار نامتقارن و دارای مشاهده‌های دور افتاده است. با توجه به خواص مطلوب این خانواده از توزیع‌ها، پژوهش‌های جدید بسیاری بر مبنای خواص این خانواده بنا شده‌اند.

تعریف ۱.۲. فرض کنید T_0 و T_1 مستقل و دارای توزیع نرمال استاندارد باشند و $T = k^{\frac{1}{2}}(U)\,|T_0|$ که در آن $k(u)$ تابعی مثبت از u است و U نیز متغیر تصادفی با تابع توزیع $H(.;v)$ و تابع چگالی $h(.;v)$ است که در آن v یک اسکالر یا بردار است. در این صورت $X \sim SMSN(\mu, \sigma^2, \lambda, H)$ و نمایش تصادفی آن به صورت

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

است (باسو و همکاران، [۱۳] [٤]). برای ساده تر شدن محاسبات $k(u) = \frac{1}{u}$ و v اسکالر در نظر گرفته می‌شود، بنابراین تابع چگالی X به صورت

$$\begin{aligned} f(x) &= 2\int_0^\infty f_{X|U}(x)dH(u;\nu) \\ &= 2\int_0^\infty \phi(x;\mu,u^{-1}\sigma^2)\Phi(\frac{u^{\frac{1}{2}}\lambda(x-\mu)}{\sigma})dH(u;v), \end{aligned}$$

نوشته می‌شود (گری و همکاران، [۱۴] [۷]). با تغییر توزیع U انواع توزیع‌های آمیخته–مقیاس چوله–نرمال حاصل می‌شوند.

تعریف ۲.۲. فرض کنید متغیر تصادفی Z دارای توزیع $SN(0,\sigma^2,\lambda)$ باشد. همچنین $k(u)$ مستقل از Z باشد. متغیر تصادفی X متعلق به خانواده توزیع‌های آمیخته–مقیاس چوله–نرمال است هرگاه

$$X \overset{d}{=} \mu + k^{\frac{1}{2}}(U)Z \qquad (۲)$$

با توجه به تعریف توزیع SN ، تعریف‌های (۱.۲) و (۲.۲) معادل هستند، حال با توجه به رابطه (۲) توزیع $X \mid U = u \sim SN(\mu, u^{-1}\sigma^2, \lambda)$ حاصل می‌شود. در این تعریف اگر $Z \sim N(0,\sigma^2)$ ، آنگاه $X \sim SMN$ [۱۵] (μ, σ^2, H) به دست می‌آید.

[۹] Cristina and et al.

[۱۰] Scale Mixtures of Skew-Normal

[۱۱] Skew t

[۱۲] Skew-Slash

[۱۳] Basso and et al.

[۱۴] Garay and et al.

[۱۵] Scale Mixtures of Normal

۳ معرفی توزیع اسلش

تعریف ۳.۱. گوییم متغیر تصادفی $X = \frac{Z}{U^{\frac{1}{\nu}}}$ دارای توزیع اسلش با پارامتر شکل ν است هرگاه $Z \sim N(0,1)$ و $U \sim U(0,1)$ مستقل از یکدیگر باشند و آن را با نماد $X \sim SL(\nu)$ نمایش می‌دهیم. تابع چگالی آن به صورت زیر است:

$$f(x;\nu) = \nu \int_0^1 t^{\nu} \phi(xt) dt \qquad -\infty < x < \infty \quad , \quad \nu > 0 \tag{۳}$$

که در آن $\phi(.)$ تابع چگالی نرمال استاندارد است.

تابع توزیع تجمعی اسلش به صورت زیر است:

$$F(x;\nu) = \nu \int_0^1 t^{\nu-1} \Phi(xt) dt, \tag{٤}$$

که در آن $\Phi(.)$ تابع توزیع نرمال استاندارد است.

تابع مولد گشتاور توزیع اسلش وجود ندارد امّا گشتاورهای آن به صورت زیر محاسبه می‌شود:

$$E(X^n) = E(Z^n).E(U^{\frac{-n}{\nu}}) = E(Z^n).\frac{\nu}{\nu - n}, \qquad \nu > n. \tag{۵}$$

در نتیجه گشتاورهای اولیه به صورت زیر به دست می‌آیند:

$$\begin{aligned} \mu_1 &= E(X) = 0, && \nu > 1 \\ \mu_2 &= E(X^2) = \frac{\nu}{\nu - 2}, && \nu > 2 \\ \mu_3 &= E(X^3) = 0, && \nu > 3 \\ \mu_4 &= E(X^4) = \frac{3\nu}{\nu - 4}, && \nu > 4. \end{aligned} \tag{٦}$$

بنابراین واریانس آن به صورت زیر است:

$$Var(X) = \frac{\nu}{\nu - 2}, \quad \nu > 2. \tag{۷}$$

ضریب چولگی آن عبارتست از:

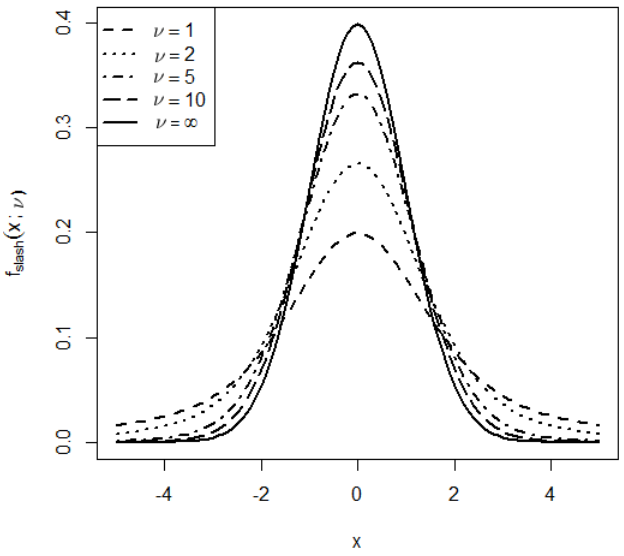
$$\delta_1(X) = 0, \tag{۸}$$

زیرا در توزیع‌های متقارن ضریب چولگی برابر صفر است.

اگر گشتاورهای X تا مرتبه ٤ موجود باشد آن‌گاه ضریب کشیدگی متغیر تصادفی اسلش X که با $\delta_2(X)$ نمایش داده می‌شود عبارت خواهد بود از:

$$\begin{aligned} \delta_2(X) &= \frac{E[(X - E(X))^4]}{[Var(X)]^3} - 3 \\ &= 3\left(\frac{1}{\nu(\nu-4)(\nu-2)^2} - 1\right), \quad \nu > 4. \end{aligned}$$

که بزرگ‌تر از صفر است و لذا چگالی اسلش کشیده است. نمودار تابع چگالی $SL(\nu)$ به ازای مقادیر مختلف ν در شکل (۱) نشان داده شده است.



شکل ۱. نمودار تابع چگالی توزیع اسلش

٤ توزیع چوله-اسلش

در این بخش به معرفی توزیع چوله-اسلش و ویژگی‌های آن می‌پردازیم.

تعریف ٤.۱. گوییم متغیر تصادفی $X = \frac{Z_\lambda}{U}$ دارای توزیع چوله-اسلش با پارامترهای ν و λ است، هرگاه $Z_\lambda \sim SN(\lambda)$ و $U \sim Beta(\nu, 1)$ مستقل از یکدیگر باشند و آن را با نماد $X \sim SSL(\nu, \lambda)$ نمایش می‌دهیم. تابع چگالی آن به صورت زیر است:

$$f(x;\nu,\lambda) = \nu \int_0^1 u^{\nu} \phi_{SN}(xu;\lambda) du, \quad x, \lambda \in \mathbb{R}, \quad \nu > 0. \tag{۹}$$

که در آن $\phi(.)$ تابع چگالی نرمال استاندارد است.

تابع مولد گشتاور توزیع چوله-اسلش وجود ندارد امّا گشتاورهای آن به صورت زیر محاسبه می‌شود:

$$E(X^n) = E(Z^n).E(U^{-n}) = E(Z^n).\frac{\nu}{\nu - n}, \qquad \nu > n. \tag{۱۰}$$

در نتیجه گشتاورهای اولیه X ، با $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$ ، به صورت زیر به دست می‌آیند:

$$\begin{aligned}
&\mu_1 = E(X) = \frac{\nu}{\nu-1}\sqrt{\frac{2}{\pi}}\delta, && \nu > 1\\
&\mu_2 = E(X^2) = \frac{\nu}{\nu-2}, && \nu > 2\\
&\mu_3 = E(X^3) = \frac{\nu}{\nu-3}\sqrt{\frac{2}{\pi}}(3\delta-\delta^3), && \nu > 3\\
&\mu_4 = E(X^4) = \frac{3\nu}{\nu-4}, && \nu > 4
\end{aligned} \qquad (11)$$

بنابراین واریانس آن

$$Var(X) = \frac{\nu}{\nu-2} - (\frac{\nu}{\nu-1})^2\frac{2}{\pi}\delta^2, \quad \nu > 2 \qquad (12)$$

و ضریب چولگی و کشیدگی آن به ترتیب برابر است با

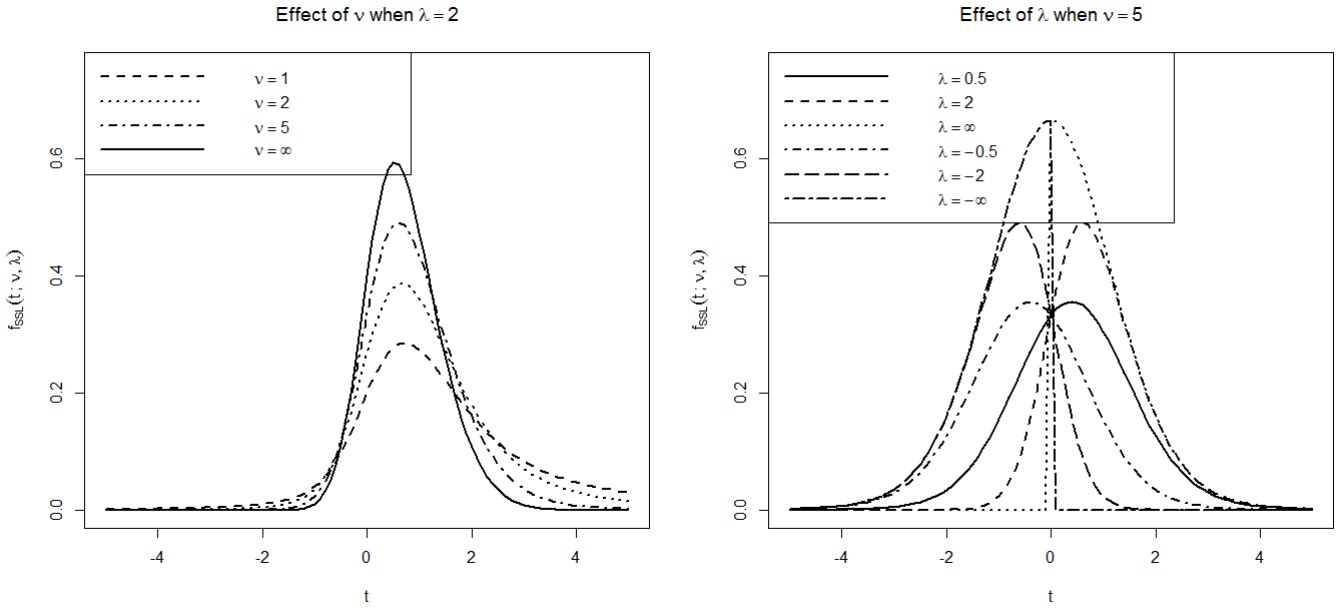
$$\delta_1(X) = \frac{\frac{4\sqrt{2}\delta^3\nu^3}{\pi^{\frac{3}{2}}(\nu-1)^3} + \frac{\sqrt{2}\nu(3\delta-\delta^3)}{\sqrt{\pi}(\nu-3)} - \frac{3\sqrt{2}\delta\nu^2}{\sqrt{\pi}(\nu-1)(\nu-2)}}{(\frac{\nu}{\nu-2} - \frac{2\delta^2\nu^2}{\pi(\nu-1)^2})^{\frac{3}{2}}}, \qquad \nu > 3 \qquad (13)$$

$$\delta_2(X) = \frac{\frac{3\nu}{\nu-4} - \frac{12\delta^4\nu^4}{\pi^2(\nu-1)^4} + \frac{12\delta^2\nu^3}{\pi(\nu-1)^2(\nu-2)} - \frac{8\delta\nu^2(3\delta-\delta^3)}{\pi(\nu-1)(\nu-3)}}{(\frac{\nu}{\nu-2} - \frac{2\delta^2\nu^2}{\pi(\nu-1)^2})^2} - 3, \qquad (14)$$

با توجه به اینکه $\lambda \in \mathbb{R}$ ، واضح است که

$$-1 < \delta = \frac{\lambda}{\sqrt{1+\lambda^2}} < 1.$$

نمودار تابع چگالی $SSL(\nu, \lambda)$ به ازای مقادیر مختلف ν و λ در شکل (۲) نشان داده شده است.



شکل ۲. نمودار تابع چگالی توزیع چوله–اسلش

همانطور که در شکل (۲) دیده می‌شود با کوچک شدن مقدار پارامتر λ، دم تابع چگالی سنگین‌تر می‌شود و با افزایش مقدار λ توزیع به نرمال بریده شده میل می‌کند. البته با کاهش مقدار λ تا $-\infty$ رفتار معکوس دارد.

توزیع SSL یک توزیع دم–سنگین است و اگر $\nu \to \infty$ آنگاه توزیع SN حاصل می‌شود.

۱.٤ نمایش تصادفی توزیع چوله–اسلش

داده‌های مالی اغلب دارای توزیعی با چولگی متوسط و کشیدگی زیاد می‌باشند. در نتیجه طبیعی است که مدل انعطاف پذیری برازش دهیم تا برازنده داده‌ها باشد. با توجه به این که در تحلیل داده‌های واقعی، مواردی نیز وجود دارند که توزیع مشاهدات متقارن نیست، لذا مطالعه بر روی تعمیم‌هایی از توزیع اسلش به حالت چوله‌ی آن مورد توجه محققان قرار گرفته است. توزیع چوله–اسلش دارای نمایش تصادفی آمیخته–مقیاسی از توزیع چوله–نرمال می‌باشد که در زیر به آن می‌پردازیم. توزیع چوله–اسلش از خانواده توزیع‌هایی

با سه پارامتر مکان، مقیاس و شکل است که در آن پارامتر شکل، چولگی توزیع را کنترل می‌کند. وانگ و جنتون [۱۵] توزیع چوله-اسلش که توزیع چوله-نرمال را تعمیم می‌دهد به صورت زیر تعریف کردند.

قضیه ۲.٤. اگر $Z \sim SN(\lambda)$ و متغیر تصادفی U که مستقل از Z است دارای توزیع بتا با تابع چگالی

$$f(u) = \nu u^{\nu-1} \qquad 0 < u < 1$$

باشد آنگاه متغیر تصادفی Y را با نمایش تصادفی $Y = \mu + \sigma U^{-1} Z$ که در آن μ پارامتر مکان، σ پارامتر مقیاس، λ پارامتر چولگی و ν پارامتر کشیدگی است و با نماد $Y \sim SSL(\mu, \sigma, \lambda, \nu)$ نمایش می‌دهیم دارای چوله-اسلش با تابع چگالی زیر است:

$$f_Y(y) = \int_0^1 2\nu u^{\nu-1} \phi(y; \mu, u^{-2}\sigma^2) \Phi\left(\frac{\lambda u (y-\mu)}{\sigma^2}\right) du, \tag{۱۵}$$

که در آن $\phi(y; \mu, u^{-2}\sigma^2)$ تابع چگالی توزیع نرمال با میانگین μ و واریانس $u^{-2}\sigma^2$ است .

فرض کنید $Y \sim SSL(\mu, \sigma, \lambda, \nu)$. امید ریاضی Y عبارتست از:

$$E(Y) = \mu + \sigma(\frac{2}{\pi})^{\frac{1}{2}} \frac{\nu}{\nu-1} (\frac{\lambda^2}{1+\lambda^2})^{\frac{1}{2}}, \qquad \nu > 1 \tag{۱٦}$$

به ازای $k \in \mathbb{N}$ ، گشتاورهای مرکزی توزیع چوله-اسلش به صورت زیر است:

$$\begin{aligned} m_k[Y] &= E\left[(Y - E(Y))^k\right] \qquad (۱۷) \\ &= 2^{\frac{(k-2)}{2}} \sigma^k \left(\frac{1}{1+\lambda^2}\right)^{\frac{k}{2}} \Sigma_{l=0}^{k} c_{lk} \lambda^l, \quad \nu > k. \end{aligned}$$

که در آن

$$c_{lk} = \Sigma_{m=0}^{l} b_{m,k-l+m,k},$$

و

$$b_{m,k-l+m,k} = (-1)^{l-m}\left[1 + (-1)^{k-l}\right] \pi^{\frac{-(l-m+2)}{2}} \times \frac{\nu}{\nu-(k-l+m)} \left(\frac{\nu}{\nu-1}\right)^{(l-m)} \binom{k}{l} \binom{l}{m} \Gamma(\frac{m+1}{2}) \Gamma(\frac{k-l+1}{2})$$

و $l \in \{,1,...,k\}$ و $m \in \{0,1,...,l\}$ ثابت هستند. در حالت خاص اگر $k-1$ عددی فرد باشد، آنگاه $c_{lk} = 0$.

با استفاده از بیان کلی که از گشتاورهای مرکزی توزیع چوله-اسلش در رابطه (۱۷) حاصل شد می‌توان واریانس و ضریب چولگی و کشیدگی توزیع چوله-اسلش را به صورت زیر به دست آورد.

واریانس توزیع چوله-اسلش به صورت زیر است:

$$V[Y] = m_2[Y] = \sigma^2 \frac{c_{02} + c_{22}\lambda^2}{1+\lambda^2},$$

که در آن $c_{02} = \frac{\nu}{\nu-2}$ و $c_{22} = \frac{\nu}{\nu-2} - \frac{2}{\pi}\left(\frac{\nu}{\nu-1}\right)^2$ که با قرار دادن $k=2$ در معادله (۱۷) واریانس توزیع چوله-اسلش به دست می‌آید.

ضریب چولگی توزیع چوله-اسلش به صورت زیر است:

$$S[Y] = \frac{m_3[Y]}{m_2[Y]^{\frac{3}{2}}} = 2^{\frac{1}{2}} \frac{c_{13}\lambda + c_{33}\lambda^3}{(c_{02} + c_{22}\lambda^2)^{\frac{3}{2}}}, \qquad \nu > 3$$

که در آن

$c_{02} = \frac{\nu}{\nu-2}$ ، $c_{22} = \frac{\nu}{\nu-2} - \frac{2}{\pi}\left(\frac{\nu}{\nu-1}\right)^2$ ،

$c_{13} = \frac{3}{\pi^{\frac{1}{2}}}\left(\frac{\nu}{\nu-3} - \frac{\nu}{\nu-2}\frac{\nu}{\nu-1}\right)$ و

$c_{33} = \frac{4}{\pi^{\frac{3}{2}}}\left(\frac{\nu}{\nu-1}\right)^3 - \frac{3}{\pi^{\frac{1}{2}}}\frac{\nu}{\nu-2}\frac{\nu}{\nu-1} + \frac{2}{\pi}\frac{\nu}{\nu-3}$ می‌باشد. با قرار دادن $k=3$ در معادله (۱۷) ضریب چولگی توزیع چوله-اسلش به دست می‌آید.

ضریب کشیدگی توزیع چوله-اسلش به صورت زیر است:

$$K[Y] = \frac{m_4[Y]}{m_2[Y]^2} = 2\frac{c_{04} + c_{24}\lambda^2 + c_{44}\lambda^4}{(c_{02} + c_{22}\lambda^2)^2}, \qquad \nu > 4$$

که در آن

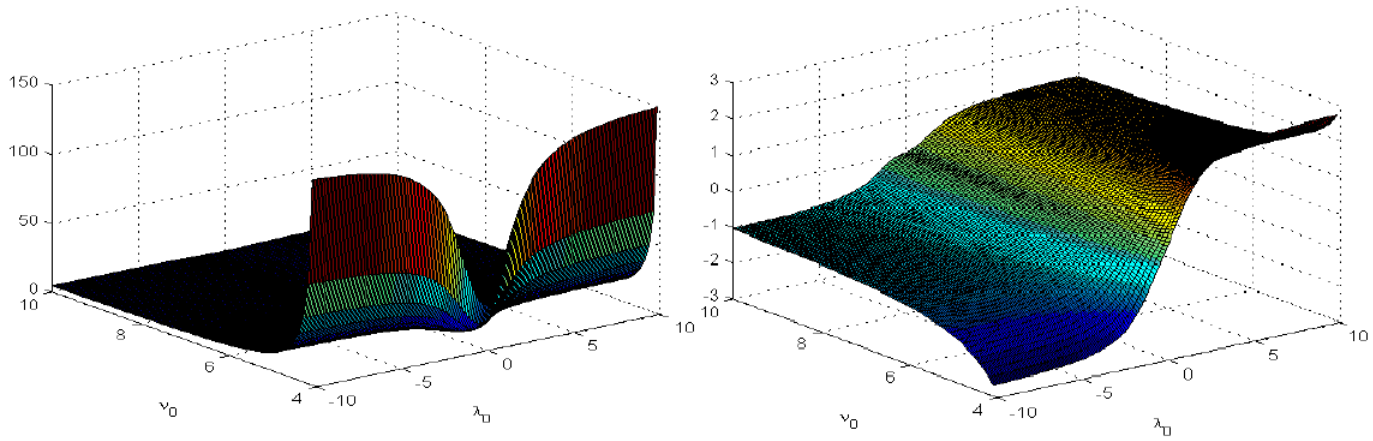
$c_{04} = \frac{3}{2}\frac{\nu}{\nu-4}$ ، $c_{02} = \frac{\nu}{\nu-2}$ ،

$c_{24} = \frac{6}{\pi}\frac{\nu}{\nu-2}(\frac{\nu}{\nu-1})^2 - \frac{12}{\pi}\frac{\nu}{\nu-3}\frac{\nu}{\nu-1} + 3\frac{\nu}{\nu-4}$ ،

$c_{44} = -\frac{6}{\pi^2}(\frac{\nu}{\nu-1})^4 + \frac{12}{\pi}\frac{\nu}{\nu-2}(\frac{\nu}{\nu-1})^2 - \frac{8}{\pi}\frac{\nu}{\nu-3}\frac{\nu}{\nu-1} + \frac{3}{2}\frac{\nu}{\nu-4}$ و

$c_{22} = \frac{\nu}{\nu-2} - \frac{2}{\pi}\left(\frac{\nu}{\nu-1}\right)^2$.

با قرار دادن $k=4$ در معادله (۱۷) ضریب کشیدگی توزیع چوله-اسلش به دست می‌آید. توزیع چوله-اسلش دارای کشیدگی و چولگی زیادی است. برای درک این مطلب شکل (۳) ضریب کشیدگی و چولگی را به ازای مقادیر λ و ν به ترتیب در فاصله (۱۰و۱۰-) و (۱۰و٤) نشان می‌دهد.



شکل۳. نمودار سمت راست ضریب چولگی توزیع چوله-اسلش و سمت چپ ضریب کشیدگی توزیع چوله-اسلش

مشاهده می‌کنیم که همانند توزیع چوله-نرمال ضریب چولگی به ازای $\lambda = 0$ برابر با صفر است و به ازای مقادیر مثبت λ ، مثبت و به ازای مقادیر منفی λ ، منفی است. ν تأثیر کمی در چولگی توزیع چوله-اسلش دارد و λ تأثیر کمی در کشیدگی توزیع چوله-اسلش دارد. از طرف دیگر با کاهش ν ضریب کشیدگی افزایش می‌یابد.

قضیه ۳.٤. اگر $Y \sim SSL(0, \sigma^2, \lambda, \nu)$ آنگاه $-Y \sim SSL(0, \sigma^2, -\lambda, \nu)$.

اثبات. داریم: $P(-Y \leq y) = 1 - P(Y \leq -y)$. در این صورت داریم $f_{-Y}(y) = f_Y(-y)$. بنابراین با توجه به رابطه (۱۵) و تقارن توزیع نرمال داریم :

$$\begin{aligned} f_{-Y}(y) &= 2\nu \int_0^1 u^{\nu-1}\phi(-y;0,u^{-1}\sigma^2)\Phi(-\frac{\lambda y u^{\frac{1}{2}}}{\sigma})du \\ &= 2\nu \int_0^1 u^{\nu-1}\phi(y;0,u^{-1}\sigma^2)\Phi(-\frac{\lambda y u^{\frac{1}{2}}}{\sigma})du \end{aligned}$$

نتیجه به دست آمده نشان دهنده توزیع $SSL(0, \sigma^2, -\lambda, \nu)$ است. □

۵ مدل رگرسیون خطی

گاهی دو یا چند متغیر تأثیر عمده ای روی متغیر وابسته دارند. در این وضعیت از رگرسیون چندگانه جهت پیش بینی متغیر وابسته استفاده می‌شود. در رگرسیون چندگانه نیز فرض خطی بودن متغیرها برقرار می‌باشد. مدل رگرسیون خطی چندگانه برای $i = 1, 2, ..., n$ به صورت

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ki} + \varepsilon_i \qquad (۱۸)$$

و یا به شکل برداری $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ وبا فرض‌های

الف) $E(\boldsymbol{\varepsilon}) = 0$

ب) $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

در نظر گرفته می‌شود. در این مدل $\mathbf{Y}$ بردار متغیر پاسخ و n بعدی است و $\mathbf{X}$ ماتریس متغیرهای مستقل و $n \times p$ بعدی است که در آن n تعداد مشاهدات و $p = k+1$ تعداد ضرایب رگرسیونی است. اولین ستون ماتریس X بردار $j = (1, 1, ..., 1)'$ است و هر یک از سایر ستون‌ها متعلق به یک متغیر مستقل و ثابت است. $\boldsymbol{\varepsilon}$ خطای مدل بنا بر فرض معمول توزیع آن نرمال است . $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)'$ ضرایب رگرسیونی و قابل برآورد به روش ماکسیمم درستنمایی و حداقل مربعات در صورت مشخص نبودن توزیع خطا هستند. تابع درستنمایی به صورت

$$L(\boldsymbol{\beta}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left\{-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

است. برآورد نااریب ضرایب رگرسیونی توسط هر دو روش به صورت $\widehat{\boldsymbol{\beta}} = (\mathbf{XX})^{-1}\mathbf{X}'\mathbf{Y}$ دست می‌آید که حاصل آن بردار p بعدی است. توزیع این برآوردگر نرمال با میانگین $\boldsymbol{\beta}$ و واریانس $\sigma^2(\mathbf{XX})^{-1}$ است. دلیل نامگذاری این مدل، خطی بودن رابطه متغیر پاسخ نسبت به ضرایب رگرسیونی است. بنابراین در معادله (۱۸) متغیرهای توضیحی می‌توانند با هر توان و یا شکل دیگری نوشته شوند. بسیاری از نویسندگان خواص مطلوب این مدل را بررسی کرده اند و در برخی موارد نیز بعضی از فرض‌های مدل نادرست در نظر گرفته شده و ویژگی‌های مطلوب یا نامطلوب حاصل مورد بررسی قرار گرفته اند. برای رفع مشکل برقرار نبودن

فرض‌ها در داده‌های واقعی راه حل‌های مناسبی توسط محققان مختلف ارائه شده است.

۱.۵ روشهای متفاوت برای برآورد ضرایب رگرسیون

یکی از مباحث اصلی تحلیل‌های رگرسیونی، برآورد پارامترهای مدل است. اگر تابع رگرسیون جامعه را با

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ki} + \varepsilon_i$$

و برآوردهای β_i و y_i را به ترتیب با $\widehat{\beta_i}$ و $\widehat{y_i}$ نشان دهیم، مدل رگرسیون برازش شده عبارت است از

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + ... + \widehat{\beta_k} x_{ki} \qquad (۱۹)$$

اختلاف بین مقادیر مشاهده شده (y_i) و مقادیر برازش شده $(\widehat{y_i})$ را به عنوان خطا در نظر می‌گیرند و با e_i نمایش می‌دهند. $\widehat{\beta_i}$ ها قابل برآورد هستند و در نتیجه می‌توان مدل رگرسیون را برآورد کرد. امّا پارامترهای واقعی جامعه هیچگاه قابل مشاهده و اندازه گیری نیستند زیرا اساساً ε_i قابل مشاهده نیست. برای برآورد مدل‌های رگرسیون، بسته به نوع مدل، روش‌های متفاوت حداقل مربعات (OLS) و حداکثر درستنمایی به کار برده می‌شود.

٦ مثال کاربردی:تحلیل داده‌های نرخ پاکت‌های سیگار

در این بخش با ارائه‌ی یک مثال تجربی به بررسی اهمیت کاربرد توزیع چوله-اسلش در مدل‌های رگرسیونی می‌پردازیم. داده‌هایی [۱۶] که در این مثال مورد بررسی قرار می‌گیرند در سال ۱۹۷۰ توسط سازمان بیمه آمریکا به منظور بررسی الگوی مصرفی سیگار، در ۵۱ ایالت آمریکا و همچنین کلمبیا جمع آوری شده‌اند (چاترجی و هادی؛ [۱۷] [۵]).

مجموعه داده‌ها شامل متغیرهای زیر است:

X_1 (Age) :میانه سن افرادی که در ایالت زندگی می‌کنند

X_2 (Hs) :نسبت افراد بالای ۲۵ سال در ایالت که دبیرستان را تمام کرده‌اند

X_3 $(Income)$: سرانه‌ی درآمد افراد در ایالت

X_4 $(Female)$: نسبت زنان ایالت

X_5 $(Sale)$: نرخ پاکت‌های سیگار فروخته شده در ایالت بر اساس قیمت سرانه

X_6 $(Black)$: نسبت افراد سیاه پوست ایالت

Y $(Price)$: قیمت متوسط وزنی یک پاکت سیگار در ایالت .

این داده‌ها تنها با استفاده از مدل رگرسیونی نرمال در متون مختلف مرتبط با مباحث رگرسیونی برازش داده شده است. ما در این مقاله با جایگزینی توزیع چوله-اسلش نشان می‌دهیم که این مدل دارای عملکرد بهتری است. بدین منظور ابتدا با فرض نرمال بودن متغیر پاسخ $(sale)$ به شرط متغیرهای توضیحی، مدل رگرسیونی

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon_i \qquad (۲۰)$$

را به روش کلاسیک بر روی داده‌ها برازش می‌دهیم که نتایج در جدول (۱) مشاهده می‌شود.

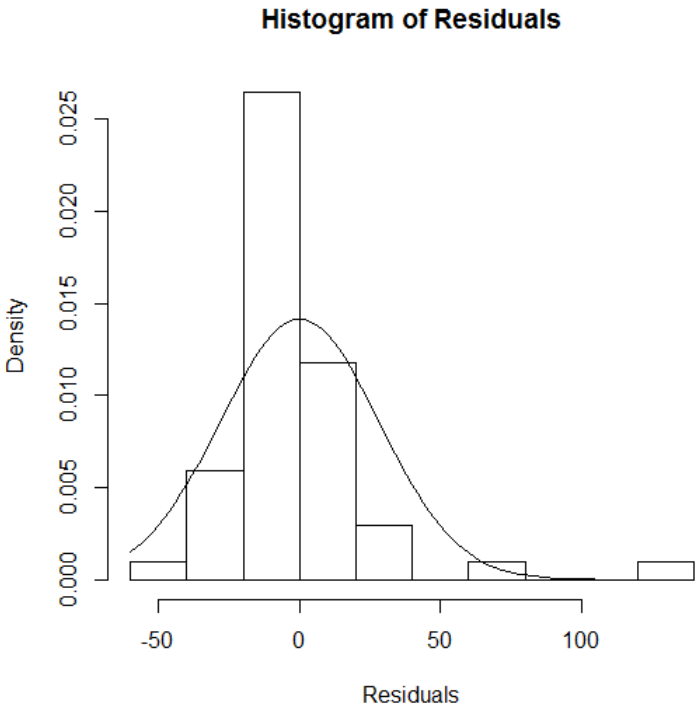
جدول ۱: تحلیل OLS پارامترهای مدل نرمال برای مدل رگرسیونی (۲۰)

پارامترها	برآورد پارامترها	انحراف معیار	$p-value$
β_0	۱۰۳/۳٤٤۸	۲٤۵/٦۰۷۱	۰/٦۷۵۹
β_1	٤/۵۲۰٤	۳/۲۱۹۷	۰/۱٦۷۳
β_2	-۰/۰٦۱٦	۰/۸۱٤٦	۰/۹٤۰۰
β_3	۰/۰۱۸۹	۰/۰۱۰۲	۰/۰۷۰۳
β_4	۰/۳۵۷۵	۰/٤۸۷۲	۰/٤٦٦۹
β_5	-۱/۰۵۲۸	۵/۵٦۱۰	۰/۸۵۰۷
β_6	-۳/۲۵٤۹	۱/۰۳۱٤	۰/۰۰۲۸
σ^2	۱۵/۱۸٦۰	۳/٦۹٦٦	۰/۰۸

برای بررسی نرمال بودن مانده‌های مدل، بافت نگار مانده‌ها رسم کرده ایم که در شکل (٤) مشاهده می‌شود.

[۱۶] داده‌ها در سایت $http://www.ilr.cornell.edu/\ hadi/RABE4$ قابل دسترسی هستند .

[۱۷] Chatterjee and Hadi



شکل٤. نمودار بافت نگار مانده‌ها

باتوجه به شکل فوق می‌توان نتیجه گرفت که مدل نرمال برای داده‌های این مثال مناسب نیست.
همچنین آزمون نرمال بودن را برای مانده‌های مدل انجام دادیم. نتایج در جدول (۲) نشان داده شده اند.

جدول ۲: آزمون شاپیرو–ویلک برای نرمال بودن مانده‌ها

$p-value$	درجه آزادی	آماره
۰	۵۱	۰/۸۳۲

با توجه به میزان $p-value$ در جدول (۲) ، فرض نرمال بودن مانده‌های مدل در سطح ۰/۰۵ رد می‌شود. همچنین نمودار مبتنی بر این آزمون در شکل (۵) رسم شده است.

شکل۵. نمودار $Q-Q$ نرمال (برای مانده‌های مدل با خطای نرمال)

نتایج فوق که حاکی از وجود کشیدگی بیشتر، سنگین‌تر بودن دم‌ها و چولگی توزیع خطای مدل نسبت به نرمال است انگیزه‌ای برای برازش مدل رگرسیونی مناسب ارائه می‌دهد. در این راستا با توجه به ویژگی‌های توزیع چوله–اسلش این توزیع را به عنوان یک جایگزین در نظر می‌گیریم که نتایج در بخش بعد بیان می‌کنیم.

۷ مقایسه توزیع چوله–اسلش، توزیع اسلش و توزیع نرمال در مدل‌های رگرسیونی چندگانه معمولی

توزیع نرمال در اکثر زمینه‌های آمار مورد استفاده است امّا در بعضی مسائل، توزیع نرمال پاسخگو نمی‌باشد و لازم است توزیع دیگری جانشین آن گردد به خصوص زمانی که توزیع داده‌ها دم سنگین تر و دارای کشیدگی بیشتر و چولگی نسبت به توزیع نرمال باشد. توزیع چوله–اسلش به دلیل داشتن ویژگی‌های مانند چولگی، کشیدگی بیشتر و دم‌های سنگین تر نسبت به توزیع نرمال، بهتر است به داده‌ها برازش شود. توزیع چوله–اسلش در دهه‌های اخیر توجه بسیاری از محققان را به خود جلب کرده است. در این بخش به معرفی مدل‌های رگرسیونی با استفاده از توزیع اسلش و چوله–اسلش می‌پردازیم. یکبار فرض می‌کنیم مؤلفه‌های خطا دارای توزیع اسلش است و برآورد حداکثر درستنمایی پارامترهای

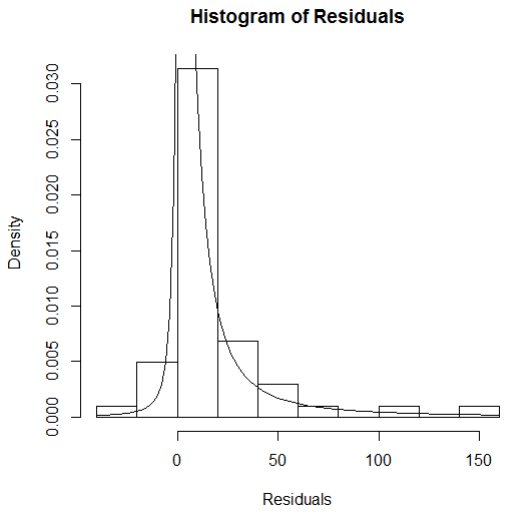
مدل را به دست می‌آوریم و بار دیگر فرض می‌کنیم مؤلفه‌های خطا دارای توزیع چوله-اسلش می‌باشد و برآورد پارامترهای آن را به دست می‌آوریم. مشاهده می‌کنیم که مدل چوله-اسلش بهترین برازش برای این مدل رگرسیونی است.

مدل رگرسیونی (۲۰) با توزیع چوله-اسلش برازش داده‌ایم که برآورد حداکثر درستنمایی پارامترهای مدل در جدول (۳) مشاهده می‌شود.

جدول ۳: برآورد پارامترهای مدل چوله-اسلش برای مدل رگرسیونی(۲۰)

پارامترها	برآورد پارامترها	انحراف معیار
β_0	-۰/۱۸۵٦	۰/۰۳۳۸
β_1	۰/۱۵۱۹	۰/۰۰۰۸
β_2	۰/۲۹۳٤	۰/۰۱٦۱
β_3	۰/۰۱۱۳	۰/۰۰۸۷
β_4	۰/۳۱۸٤	۱/۲۰٦۳
β_5	۰/۰٦٦۷	۰/۰۰۱٦
β_6	۰/۰۲٥۰	۰/۰۰٦۳
σ	۰/۸۰٥۸	۳/٥۳۷۸
ν	۱/۸۸۱۹	۰/٥۹۲٥
λ	۱/۸۸۲۰	۳/۱٥۲٥

نمودار بافت نگار مانده‌های توزیع چوله-اسلش رسم کرده‌ایم که در شکل (٦) مشاهده می‌شود.



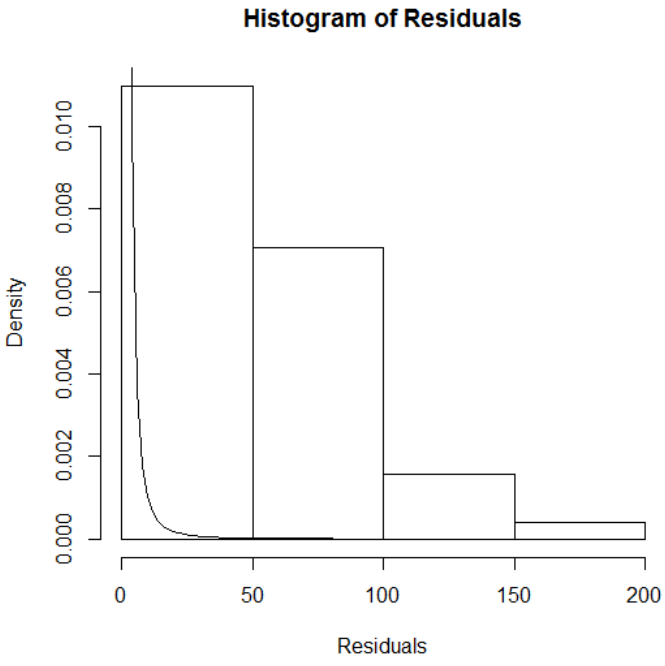
شکل٦. نمودار بافت نگار مانده‌های توزیع چوله-اسلش

مدل رگرسیونی (۲۰) با توزیع اسلش برازش داده‌ایم که برآورد حداکثر درستنمایی پارامترهای مدل در جدول (٤) مشاهده می‌شود.

جدول ٤: برآورد پارامترهای مدل اسلش برای مدل رگرسیونی(۲۰)

پارامترها	برآورد پارامترها	انحراف معیار
β_0	-۰/۰۸۸۰	۰/۰۳۳۸
β_1	۰/۰٦٦۷	۰/۰۰۰۸
β_2	۰/۱٥٤٤	۰/۰۱٥۸
β_3	۰/۰۱٥۷	۰/۰۰۰۱
β_4	۰/۰٦۱٤	۰/۰۰۰۱
β_5	-۰/۰۰۰۸	۰/۰۰۰۱
β_6	-۰/۰۹٥۲	۰/۰۰۰۳
σ	۱/٤۸۳٥	۰/۰۳۳۸
ν	۰/٦۱٦۰	۰/۰۳۳۹

نمودار بافت نگار مانده‌های توزیع اسلش رسم کرده ایم که در شکل (۷) مشاهده می شود.



شکل۷. نمودار بافت نگار مانده‌های توزیع اسلش

با استفاده از دو معیار مقایسه مدل، معیار اطلاع آکائیک و اطلاع بیزی، به مقایسه دو مدل اسلش و چوله-اسلش پرداختیم که نتایج در جدول ٥ نشان داده شده اند.

جدول ۵: مقادیر معیار اطلاع آکائیک و اطلاع بیزی برای دو مدل اسلش و چوله–اسلش

	مدل چوله–اسلش	مدل اسلش
AIC	٥٤٣/٩٦٦٥	٥٥٣/٨٥٧٩
BIC	٥٦٠/١٤٤٣	٥٧١/٢٤٤٣

از بین دو اسلش و چوله–اسلش، توزیع چوله–اسلش دارای کمترین مقدار معیار اطلاع آکائیک (AIC) و معیار اطلاع بیزی (BIC) است. پس نسبت به توزیع اسلش بهترین برازش برای مدل رگرسیونی رابطه (۲۰) است.

۸ نتیجه گیری

در این مقاله به بررسی توزیع‌های اسلش و چوله–اسلش و ویژگی‌های آنها پرداختیم. در برازش مدل‌های رگرسیونی چندگانه معمولی با توزیع چوله–اسلش و اسلش برازش دادیم و با استفاده از داده‌های واقعی نشان دادیم که مدل چوله–اسلش جایگزین مناسبی برای توزیع نرمال است. برای مقایسه دو مدل از معیارهای مختلف استفاده شد و مشاهده کردیم که توزیع چوله–اسلش نسبت به دو مدل نرمال و اسلش در بعضی حالات برای مؤلفه‌های خطا مناسب‌تر است.

مراجع

[1] Arslan, O. (2008). An alternative multivariate skew-slash distribution, *Statistics and Probability Letters*,**78**, 2756–2761.

[2] Azzalini, A. (1985). A class of distribution which includes the normal ones, *Scandinavian Journal of Statistics* ,**12**, 171-178.

[3] Azzalini, A. and Dalla-Valle, A. (1996). The multivariate skew-normal distribution, *Biometrika* , **83**, 715–726.

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

[5] Chatterjee, S. and Hadi, A. (2006). *Regression analysis by example*. John Wiley and Sons,Inc.

[6] Cristina, G. and Fuente, D. L and Galeano, P. and Michael P. W. (2012). Modeling financial time series with the skew slash distribution. *Statistics and Econometrics*, **126**, 1-26.

[7] Garay, A. M and Lachos, V. H and 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.

[8] Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications*. A Journey Beyond Normality, Chapman and Hall/CRC Press, Boca. Raton, Fla.

[9] Genton, M. G. and Loperfido, N. (2005). Generalized skew-elliptical distributions and their quadratic forms, *Ann. Inst Statist Math*; to appear.

[10] Jamshidian, M. (2001). A note on parameter and standard error estimation in adaptive robust regression. *Journal of Statistical Computation and Simulation*, **71**, 11–27

[11] Kafadar, K. (1982). A biweight approach to the one-sample problem. *Journal of the American Statistical Association*, **77**, 416–424.

[12] Kashid, D. N and Kulkarni, S. R. (2003). Subset selection in multiple linear regression with heavy tailed error distribution.*Journal of Statistical Computation and Simulation*, **73**, 791–805

[13] Morgenthaler, S. (1986). Robust confidence intervals for a location parameter: the configural approach. *Journal of the American Statistical Associatin*, **81**, 518–525.

[14] Rogers, W. H. and Tukey, J. W. (1972). Understanding some long-tailed symmetrical distributions. *Statistica Neerlandica*, **26**, 211–226.

[15] Telford, R . D and cunningham, R . B. (1991). Sex, sport and body-size dependency of hematology in highly trained athletes. *Medicine and science in sports and Exercise*, **23**, 788-794.

[16] Wang, J. and Genton, M. G. (2006). The multivariate skew-slash distribution, *Journal of Statistical Planning and Inference*, **136**, 209–220.