استفاده از توزیع اسلش بجای توزیع نرمال در مدل‌های رگرسیونی چندگانه

زهرا نیکنام
دانشکده آمار، دانشگاه اصفهان

چکیده

در بسیاری از تحلیل داده‌های آماری فرض متداول نرمال بودن توزیع مشاهدات است. امّا نقض این فرض در تحلیل داده‌های واقعی امکان پذیر است. به همین دلیل کلاس‌های جدیدی از توزیع‌ها که برای داده‌های مختلف دارای انعطاف‌پذیری بیشتری هستند، به عنوان جایگزین توزیع نرمال پیشنهاد شده است. در این رابطه می‌توان به توزیع اسلش که یک توزیع دم-سنگین و متقارن است، اشاره کرد که در دهه‌ی اخیر از سوی پژوهشگران زیادی مورد توجه قرار گرفته است. در این مقاله با استفاده از مجموعه داده‌های واقعی که در متون مربوط به مبحث رگرسیون با توزیع نرمال برازش داده شده است ما با در نظر گرفتن مؤلفه‌های خطا با توزیع اسلش به اهمیت این توزیع در مباحث رگرسیونی می‌پردازیم.

واژه‌های کلیدی: توزیع آمیخته-مقیاس نرمال، توزیع دم سنگین، توزیع اسلش، رگرسیون.

۱ مقدمه

بسیاری از استنباطهای آماری با فرض نرمال بودن داده‌ها صورت می‌گیرد. در صورتی که مشاهدات تأثیرگذاری در تحلیل داده‌ها موجود نباشند که فرض نرمال بودن را خدشه‌دار سازند این کاربردها به نتایج معتبری منجر می‌شوند. محققان تحقیقات زیادی جهت جایگزین کردن توزیع‌های متفاوتی به جای نرمال انجام داده اند که از آن جمله می‌توان به توزیع اسلش [۱] اشاره کرد. توزیع اسلش که علاوه بر داشتن دم‌های سنگین‌تر، نسبت به نرمال از کشیدگی بیشتری نیز برخوردار است، در استنباطهای استوار [۲] دارای اهمیت است. توزیع اسلش همانند توزیع

[۱] Slash distribution

[۲] Robust

نرمال استاندارد حول مبدأ متقارن است، امّا از دمهای سنگینی نسبت به آن برخوردار است که همین ویژگی سبب وجود مقادیری در میان مشاهدات نمونه می‌شود که دور افتاده بودن آنها نسبت به مشاهدات حاصل از توزیع نرمال استاندارد کاملاً واضح است. توزیع نرمال استاندارد برای حالتی که نمونه تصادفی فاقد مقادیر دور افتاده باشد و توزیع اسلش برای حالتی که نمونه تصادفی ازمقادیردورافتاده برخوردار باشد مناسبند. تعریف‌های متعدی از توزیع اسلش یک متغیره و چندمتغیره توسط محققان صورت گرفته است. توزیع اسلش و بررسی ویژگی‌های آن توسط محققان زیادی از جمله روگرز و توکی [۳] [۱۶]، مورجنتالر [۴] [۱۴]، جمشیدیان [۵] [۹]، کاشید و کالکارنی [۶] [۱۲] صورت گرفته است. کفادار [۷] [۱۱] برآورد حداکثر درستنمایی پارامترهای مقیاسی و مکانی را برای توزیع اسلش استاندارد به دست آورد. بسط توزیع اسلش یک متغیره به حالت چند متغیره اولین بار توسط لانگ و سینشیمر [۸] [۱۳] انجام شد. جنس [۹] [۵] توزیع اسلش تعمیم یافته یک متغیره را با استفاده از آمیخته-مقیاس توزیع نمایی-توانی به دست آورد و خصوصیات آن را مورد بررسی قرار داد. گومز و همکاران [۱۰] [۷] خانواده جدیدی از توزیع های اسلش یک متغیره و چند متغیره که ساختار آنها بر اساس توزیع های بیضوی است را معرفی کردند. اخیراً ارسلان و جنس [۳] به معرفی توزیع اسلش یک متغیره و چند متغیره پرداخته و ویژگی‌های آن را مورد بررسی قرار داده‌اند. جنس [۶] تعمیمی از توزیع اسلش را از طریق توزیع نرمال-بتا معرفی کرد و ضرورت عملی آن را با مجموعه ای از داده‌های واقعی با در نظرگرفتن روش برآورد حداکثر درستنمایی تشریح کرد. گومز و همکاران [۸] توزیع اسلش اصلاح شده را معرفی و ویژگی‌های آن را مورد بررسی قرار دادند. با توجه به اینکه توزیع اسلش به عنوان آمیخته مقیاسی از توزیع نرمال می‌باشد ابتدا در بخش ۲ به معرفی خانواده توزیع‌های آمیخته-مقیاس نرمال می‌پردازیم و دربخش ۳ به بررسی توزیع اسلش و ویژگی های آن می‌پردازیم و دربخش ۴ مدل رگرسیون خطی چندگانه با خطای نرمال مرور می‌کنیم و دربخش ۵ به معرفی مدل رگرسیون خطی چندگانه با خطای اسلش می‌پردازیم و در پایان نیز اهمیت کاربرد این توزیع را با یک مثال تجربی نشان می‌دهیم.

۲ خانواده توزیع‌های آمیخته-مقیاس نرمال (SMN)

توزیع‌های آمیخته-مقیاس نرمال [۱۱] که نقش بسیار مهمی در مدل‌سازی آماری دارند داده های دور افتاده را پوشش می‌دهند و دم‌هایی سنگین تر از نرمال دارند. مدل‌های آمیخته-مقیاس شامل توزیع‌های تی-استیودنت، اسلش و واریانس-گاما و توزیع پیرسون نوع هفتم است.

توزیع‌های آمیخته-مقیاس نرمال توسط آمیختن یک متغیر تصادفی نرمال، Z با یك متغیر تصادفی نامنفی مقیاسی λ به صورت زیر حاصل می‌شوند:

$$X = \mu + k^{\frac{1}{2}}(\lambda)Z,$$

[۳] Rogers and Tukey
[۴] Morgenthaler
[۵] Jamshidian
[۶] Kashid and Kulkarni
[۷] Kafadar
[۸] Lang and Sinsheimer
[۹] Genç
[۱۰] Gomez and et al.
[۱۱] Scale Mixtures of Normal

که در آن μ پارامتر مکان و λ متغیر تصادفی مثبت آمیخته با تابع چگالی احتمال $h(\lambda \mid v)$ است که مستقل از $Z \sim N(۰, \sigma^۲)$ فرض می‌شود. v یك اسکالر یا بردار پارامترهای توزیع λ و $k(.)$ یک تابع وزنی مثبت است. اگر $k(\lambda) = \frac{۱}{\lambda}$ ، آنگاه داریم :

$$X \mid \lambda \sim N(\mu, \lambda^{-۱}\sigma^۲),$$

و در این صورت تابع چگالی احتمال X به صورت زیر است:

$$f(x \mid \mu, \sigma^۲, v) = \int_۰^\infty N(x \mid \mu, \lambda^{-۱}\sigma^۲)h(\lambda \mid v)d\lambda. \quad (۱)$$

با انتخاب مناسب چگالی آمیخته $h(. \mid v)$ ، کلاس توانمندی از توزیع‌های پیوسته تک مدی و متقارن با استفاده از چگالی بیان شده در (۱) به دست می‌آید که توزیع‌هایی دم-کلفت تر از نرمال تولید می‌کنند. اگر با احتمال یک، $\lambda = ۱$ (متغیر تصادفی تباهیده) باشد، آنگاه حاصل یک توزیع نرمال است.
امید ریاضی و واریانس این توزیع‌ها را با استفاده از قوانین امید ریاضی مکرر و تفکیک واریانس و با توجه به اینکه $X \mid \lambda \sim N(\mu, k(\lambda)\sigma^۲)$ در صورت وجود به صورت زیر محاسبه می نماییم:

$$\begin{aligned} E(X) &= E(E(X \mid \lambda)) \\ &= E(\mu) = \mu. \end{aligned} \quad (۲)$$

اگر $E(k(\lambda)) < \infty$ باشد آنگاه واریانس X به صورت زیر وجود دارد:

$$\begin{aligned} Var(X) &= Var(E(X \mid \lambda)) + E(Var(X \mid \lambda)) \\ &= Var(\mu) + E(k(\lambda)\sigma^۲) = \sigma^۲ E(k(\lambda)). \end{aligned}$$

میانگین همه ی توزیع‌های متعلق به این خانواده برابر μ است ولی چون پارامتر مقیاس تصادفی است واریانس‌ها متفاوت است. (آندره و مالوز [۱۲] [۲])

۳ توزیع اسلش

توزیع اسلش به عنوان توزیع متقارن و انعطاف‌پذیر نقش مهمی در مطالعات استوار دارد. توزیع اسلش دم هایی سنگین تر از توزیع نرمال استاندارد دارد اما دم‌های آن شبیه به توزیع کوشی است که این امر در شکل ۱ به وضوح دیده می شود.

توزیع اسلش و بررسی ویژگی‌های آن توسط محققان زیادی از جمله راجرز و توکی [۱۶] ، مورجنتالر [۱۴] ، جمشیدیان [۹] ، کاشید و کالکارنی [۱۲] صورت گرفته است. سپس کفادار [۱۱] برآورد حداکثر درستنمایی پارامترهای مقیاسی و مکانی برای توزیع اسلش استاندارد به دست آورد.

توزیع اسلش که به عنوان آمیخته مقیاسی از توزیع نرمال می‌باشد به صورت زیر تعریف می‌شود.

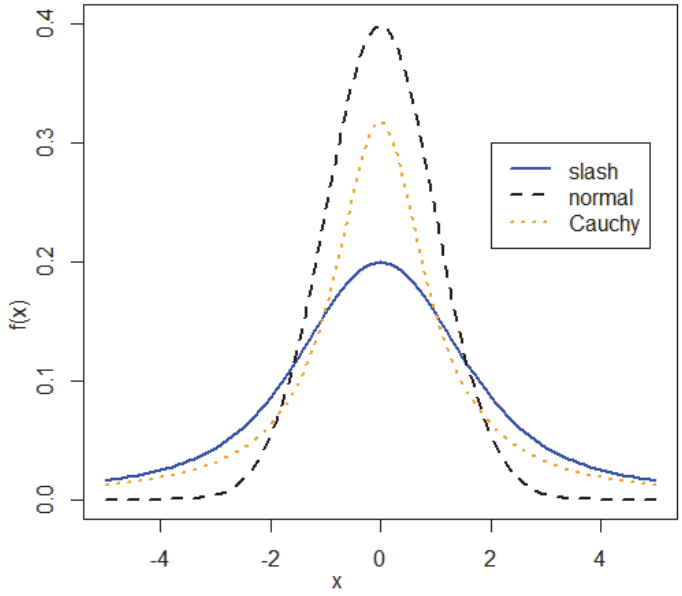
تعریف ۱. *متغیر تصادفی X دارای توزیع اسلش است و می‌نویسیم*
$X \sim SL(\mu, \sigma^۲, q)$ *هرگاه نمایش تصادفی آن به صورت*

$$X = \mu + \sigma\lambda^{-۱}Z,$$

باشد، که در آن $Z \sim N(۰, ۱)$ و $\lambda \sim Beta(q, ۱)$ مستقل از یکدیگرند.

تابع چگالی احتمال توزیع اسلش که به عنوان آمیخته-

[۱۲] Andrew and Mallows



شکل ۱: نمودار تابع چگالی توزیع های اسلش و نرمال و کوشی

مقیاسی از توزیع نرمال است، به صورت زیر است :

$$f(x \mid \mu, \sigma^2, q) = q \int_0^1 \lambda^{q-1} N(x \mid \mu, \frac{\sigma^2}{\lambda}) d\lambda,$$

که در آن چگالی λ به فرم زیر می باشد:

$$h(\lambda \mid q) = q\lambda^{q-1} I_{(0,1)}.$$

بنابراین، توزیع اسلش معادل با شکل سلسله مراتبی زیر است :

$$\lambda \mid q \sim Beta(q, 1),$$

و

$$X \mid \mu, \sigma^2, q, \lambda \sim N(\mu, \frac{\sigma^2}{\lambda}),$$

که در آن $Beta(.,.)$ نماد توزیع بتا می‌باشد. برای به دست آوردن تابع چگالی احتمال توزیع اسلش به صورت زیر عمل می کنیم:

$$f(x \mid \mu, \sigma^2, q) = q \int_0^1 \frac{1}{\sqrt{2\pi\frac{\sigma^2}{\lambda}}} \exp\left\{\frac{\lambda(x-\mu)^2}{2\sigma^2}\right\} \lambda^{q-1} d\lambda$$

$$= \begin{cases} \frac{q}{\sqrt{2\pi\sigma^2}} \frac{2^{(q+\frac{1}{2})}}{s^{(q+\frac{1}{2})}} \Gamma(q+\frac{1}{2}, \frac{s}{2}), & x \neq \mu, \\ \frac{2q}{(2q+1)\sqrt{2\pi\sigma^2}}, & x = \mu, \end{cases} \quad (3)$$

که در آن

$$z = \frac{(x-\mu)}{\sigma},$$

و

$$\Gamma(a, z) = \int_0^z u^{a-1} e^{-u} du,$$

تابع گامای ناقص است. بنابراین داریم:

$$f(x \mid \mu, \sigma^2, q) = \begin{cases} \frac{q}{\sqrt{2\pi\sigma^2}} \frac{2^{(q+\frac{1}{2})}}{z^{(2q+1)}} \Gamma(q+\frac{1}{2}, \frac{z^2}{2}), & z \neq 0, \\ \frac{2q}{(2q+1)\sqrt{2\pi\sigma^2}}, & z = 0. \end{cases}$$

میانگین، نما و واریانس توزیع اسلش که به عنوان

آمیخته-مقیاسی از توزیع نرمال است، عبارتست از :

$$E(X) = Mode(X) = \mu,$$

$$Var(X) = \frac{q}{q-۱}\sigma^۲,\ q > ۱.$$

تذکر ۱. *به ازای $\mu = ۰$ و $\sigma = ۱$ در رابطه (۳) توزیع اسلش استاندارد به دست می‌آید که به صورت زیر تعریف می شود.*

تعریف ۲. *گوییم متغیر تصادفی $X = \frac{Z}{U^{\frac{۱}{q}}}$ دارای توزیع اسلش با پارامتر شکل q است هرگاه $Z \sim N(۰,۱)$ و $U \sim U(۰,۱)$ مستقل از یکدیگر باشند و آن را با نماد $X \sim SL(q)$ نمایش می‌دهیم. تابع چگالی آن به صورت زیر است:*

$$f(x;q) = q\int_۰^۱ t^q\phi(xt)dt;\ -\infty < x < \infty,\ q > ۰ \quad (۴)$$

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

تذکر ۲. *به ازای $\mu = ۰$ و $\sigma = ۱$ ، $q = ۱$ در رابطه (۳) توزیع اسلش کانونی به دست می‌آید (جانسون [۱۳] و همکاران [۱۰]) که تابع چگالی آن به صورت زیر است:*

$$f(x,۱) = \begin{cases} \frac{۱}{\sqrt{۲\pi}}(۱ - e^{-\frac{x^۲}{۲}}), & x \neq ۰, \\ \\ \frac{۱}{۲\sqrt{۲\pi}}, & x = ۰. \end{cases}$$

تابع توزیع تجمعی اسلش به صورت زیر است:

$$F(x;q) = q\int_۰^۱ t^{q-۱}\Phi(xt)dt, \quad (۵)$$

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

[۱۳] Johnson

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

$$\begin{aligned} E(X^n) &= E(Z^n).E(U^{\frac{-n}{q}}) \\ &= E(Z^n).\frac{q}{q-n},\ q > n. \end{aligned} \quad (۶)$$

در نتیجه گشتاورهای اولیه به صورت زیر به دست می‌آیند:

$$\begin{aligned} \mu_۱ &= E(X) = ۰, && q > ۱, \\ \mu_۲ &= E(X^۲) = \frac{q}{q-۲}, && q > ۲, \\ \mu_۳ &= E(X^۳) = ۰, && q > ۳, \\ \mu_۴ &= E(X^۴) = \frac{۳q}{q-۴}, && q > ۴. \end{aligned} \quad (۷)$$

بنابراین واریانس آن به صورت زیر است:

$$Var(X) = \frac{q}{q-۲}, \quad q > ۲. \quad (۸)$$

ضریب چولگی آن عبارتست از:

$$\delta_۱(X) = ۰, \quad (۹)$$

در توزیع‌های متقارن ضریب چولگی برابر صفر است.

اگر گشتاورهای X تا مرتبه ۴ موجود باشد آنگاه ضریب کشیدگی متغیر تصادفی اسلش X که با $\delta_۲(X)$ نمایش داده می‌شود عبارت خواهد بود از:

$$\begin{aligned} \delta_۲(X) &= \frac{E[(X-E(X))^۴]}{[Var(X)]^۳} - ۳ \\ &= ۳\left(\frac{۱}{q(q-۴)(q-۲)^۲} - ۱\right), \quad q > ۴, \end{aligned}$$

که بزرگ‌تر از صفر است و لذا چگالی اسلش کشیده است.

توزیع اسلش دم کلفت تر از نرمال است و در صورتی که $q \to \infty$ به نرمال میل می‌کند. نمودار تابع چگالی توزیع اسلش به ازای مقادیر مختلف q در شکل ۲ نشان داده شده است.

۴ مدل رگرسیون چندگانه با خطای نرمال

در این بخش مدل رگرسیون با خطای توزیع نرمال را مرور می‌کنیم. مدل رگرسیون

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad (۱۰)$$

و با فرض‌های

الف) $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ یا $E(\boldsymbol{\varepsilon}) = ۰$،

ب) $cov(\mathbf{Y}) = \sigma^۲\mathbf{I}$ یا $cov(\boldsymbol{\varepsilon}) = \sigma^۲\mathbf{I}$،

را در نظر بگیرید. در این مدل $\mathbf{Y}$ بردار تصادفی $n \times ۱$ از مشاهدات، $\mathbf{X}$ یک ماتریس غیرتصادفی $n \times p$ با رتبه p ، که در آن n تعداد مشاهدات و $p = k + ۱$ تعداد ضرایب رگرسیونی است. $\boldsymbol{\beta}$ بردار پارامترهای مجهول رگرسیون $p \times ۱$ است. $\boldsymbol{\varepsilon}$ خطای مدل بنا بر فرض معمول توزیع آن نرمال است. توزیع این برآوردگر نرمال با میانگین $\boldsymbol{\beta}$ و واریانس $\sigma^۲(\mathbf{XX})^{-۱}$ است. $\boldsymbol{\beta} = (\beta_۱, \beta_۲, ..., \beta_k)'$ ضرایب رگرسیونی و قابل برآورد به روش حداکثر درستنمایی و حداقل مربعات در صورت مشخص نبودن توزیع خطا هستند. تابع درستنمایی به صورت

$$L(\boldsymbol{\beta}) = \frac{۱}{(۲\pi\sigma^۲)^{\frac{n}{۲}}} \exp\left\{-\frac{۱}{۲\sigma^۲}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right\},$$

است. برآورد نااریب ضرایب رگرسیونی توسط هر دو روش به صورت زیر به دست می آید:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{XX})^{-۱}\mathbf{X}'\mathbf{Y}, \qquad (۱۱)$$

که حاصل آن بردار p بعدی است. حال توزیع توأم و بردار باقی‌مانده $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ ، $(\widehat{\epsilon})$ را در نظر می‌گیریم. باتوجه به روابط (۱۰) و (۱۱) داریم:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}' \\ \mathbf{I} - (\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}' \end{bmatrix} \mathbf{Y} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}' \\ \mathbf{I} - (\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}' \end{bmatrix} \widehat{\boldsymbol{\epsilon}},$$

که در آن $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \sigma^۲\mathbf{I})$.

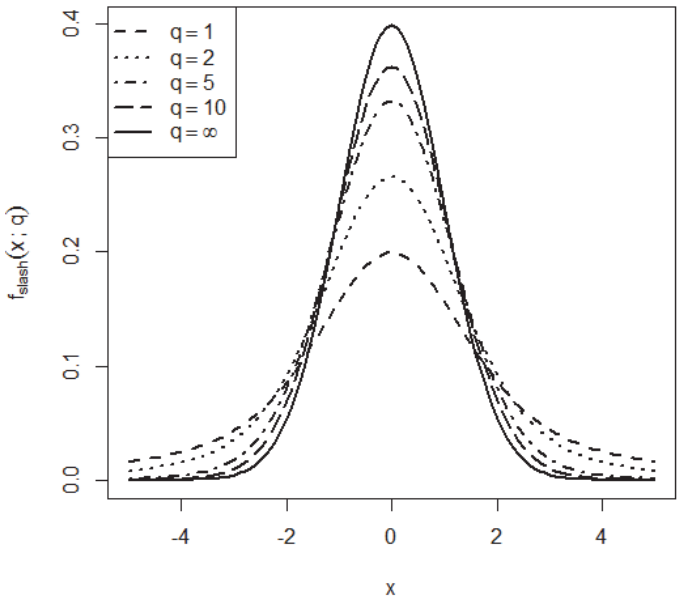
حال از آنجا که $\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\epsilon}} \end{bmatrix} - \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix}$ یک تابع خطی از $\boldsymbol{\epsilon}$ است و داریم:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\epsilon}} \end{bmatrix} - \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix} \widehat{\boldsymbol{\epsilon}} \sim N_{n+p}(\mathbf{0}, \boldsymbol{\Sigma}),$$

که در آن

$$\boldsymbol{\Sigma} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-۱} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - (\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}' \end{bmatrix}.$$

لذا به کمک روش کمترین مربعات نمی توانیم تابعی از y ها و x های نمونه برای برآورد مینیمم $\sigma^۲$ بسازیم. با وجود این، می‌توانیم یک برآوردگر نااریب برای $\sigma^۲$ بر مبنای برآوردگر کمترین مربعات $\widehat{\boldsymbol{\beta}}$ بسازیم. با فرض برقراری شرط (ب)، $\sigma^۲$ برای هر y_i ، $i = ۱, ۲, ..., n$ ثابت است.



شکل ۲: نمودار تابع چگالی توزیع اسلش به ازای مقادیرمختلف پارامتر q

σ^2 به صورت $\sigma^2 = E[y_i - E(y_i)]^2$ تعریف می‌شود و با توجه به فرض (الف) داریم:
$E(y_i) = \beta_0 x_{i1} + \beta_1 x_{i2} + ... + \beta_k x_{ik}$ که در آن x_i' سطر i ام ماتریس X است. پس σ^2 به صورت زیر نوشته می شود:

$$\sigma^2 = E[y_i - \mathbf{x}_i'\boldsymbol{\beta}]^2.$$

بنابراین می توانیم مقدار σ^2 را به صورت یک مقدار متوسط از نمونه به فرم زیر برآورد می‌کنیم.

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})^2, \qquad (۱۲)$$

که در آن n حجم نمونه و k تعداد x هاست. رابطه (۱۱) را می‌توان به صورت زیر نوشت:

$$\begin{aligned} s^2 &= \frac{1}{n-k-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \frac{\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n-k-1} = \frac{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}}{n-k-1} = \frac{SSE}{n-k-1}. \end{aligned}$$

برآوردگر s^2 برای σ^2 نااریب است. (رنچر و اسکالج [۱۴] [۱۵])

برای اولین بار فیشر [۱۵] [۴] به نتایج نادرست استفاده از توزیع نرمال به عنوان توزیع خطاهای تصادفی در مدل‌های خطی اشاره کرد.

[۱۴] Rencher and Schaalje

[۱۵] Fisher

۵ مدل رگرسیون چندگانه با خطای اسلش

فرض کنید $\epsilon_n, ..., \epsilon_۱$ متغیرهای تصادفی مستقل باشند که در آن
$\epsilon_i \sim SL(۰, \sigma^۲)$ خطای مدل رگرسیون خطی با خطای توزیع اسلش با پارامتر q است. برای برآورد بردار پارامتر $\boldsymbol{\theta} = (\boldsymbol{\beta}', q)'$ از روش حداکثر درستنمایی استفاده می کنیم. تابع لگاریتم درستنمایی توزیع اسلش به صورت زیر است:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= nlogq - nlog(۲\pi\sigma^۲) + (q + \frac{۱}{۲})log۲ \\ &+ nlog۲q - nlog(۲q + ۱) + \sum_i a(z_i), \end{aligned}$$

که در آن $a(z_i) = log\Gamma(q+\frac{۱}{۲}, \frac{z_i^۲}{۲}) - (۲q+۱)logz_i$ است. معادلات درستنمایی توزیع اسلش فرم بسته ساده‌ای ندارد، لذا این معادلات جواب صریحی برای پارامترهای آن ارائه نمی‌دهند. بنابراین باید برآوردهای حداکثر درستنمایی آن را به روشهای عددی شبیه روش نیوتن-رافسون محاسبه کرد.

۶ مثال کاربردی: تحلیل داده های جرم و جنایت

هفت متغیر زیر مربوط به داده های جرم و جنایت ۴۴ محله دنور [۱۶] می باشد.
$X_۱$:کل جمعیت (بر حسب هزار)
$X_۲$:درصد تغییر در جمعیت طی چند سال گذشته
$X_۳$: درصد کودکان (زیر ۱۸ سال) در جامعه
$X_۴$:شرکت در صرف ناهار مدارس آزاد
$X_۵$:تغییر در درآمد خانوار طی چند سال گذشته
$X_۶$:نرخ جرم و جنایت در هر ۱۰۰۰ نفر جمعیت
Y:تغییر در نرخ جرم و جنایت در طی چند سال گذشته

این داده‌ها تنها با استفاده از مدل رگرسیونی نرمال در متون مختلف بررسی شده است.

در این مقاله با در نظر گرفتن مؤلفه‌های خطا با توزیع اسلش و برآورد حداکثر درستنمایی پارامترهای مدل و با استفاده از معیارهای انتخاب، مدل مناسب برای برازش به این مدل رگرسیونی را انتخاب می‌کنیم. بدین منظور ابتدا با فرض نرمال بودن متغیر پاسخ (Y) به شرط متغیرهای توضیحی $(\mathbf{X_i})$ ، مدل رگرسیونی

$$\begin{aligned} Y_i &= \beta_۰ + \beta_۱X_۱ + \beta_۲X_۲ + \beta_۳X_۳ \\ &+ \beta_۴X_۴ + \beta_۵X_۵ + \beta_۶X_۶ + \epsilon_i, \quad (۱۳) \end{aligned}$$

را به روش کلاسیک به داده‌ها برازش می‌دهیم که نتایج در جدول ۱ مشاهده می‌شود.

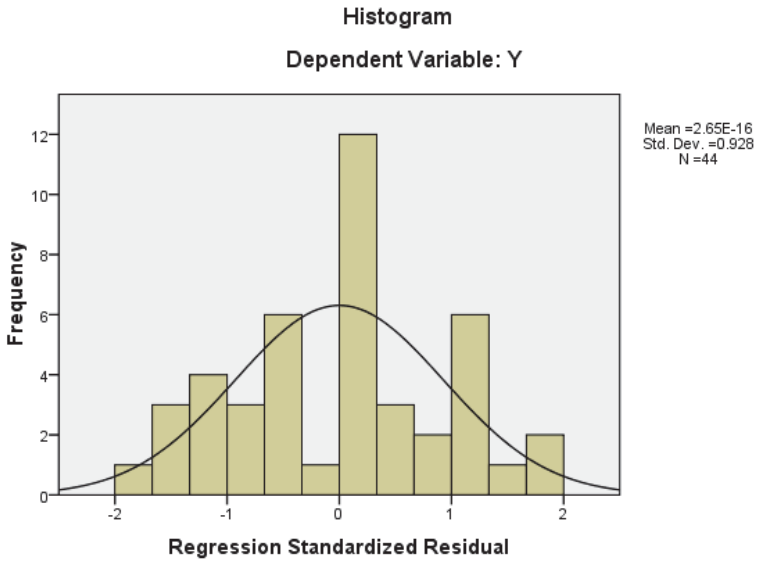
محاسبات ارائه شده در جدول ۱ این نتیجه را می‌دهد که با افزایش افراد جامعه میزان جنایت‌های رخ داده شده کمتر می شود در صورتی که این امر غیر معقول به نظر می رسد. بنابراین علت چنین نتیجه گیری برازش مدل اشتباه به داده‌های این مثال است.

برای بررسی نرمال بودن و همگنی واریانس مانده‌های مدل، نمودار بافت‌نگار باقی مانده‌ها و نمودار مانده‌ها در برابر مقادیر برازش شده در شکل های ۳ و ۴ ارائه شده است.

[۱۶] Denver Neighborhoods

جدول ۱: تحلیل OLSپارامترهای مدل نرمال برای مدل رگرسیونی (۱۳)

پارامترها	برآورد پارامترها	انحراف معیار	$p-value$
β_0	−۱۶/۸۵۶	۱۲/۶۰	۰/۱۸۹
β_1	−۰/۲۰۳	۰/۲۱۲	۰/۳۴۳
β_2	۱/۴۳۳	۰/۳۶۷	۰/۰۰۰
β_3	−۰/۴۶۸	۰/۱۳۳	۰/۰۰۱
β_4	−۰/۳۱۸	۰/۳۹۰	۰/۴۲۰
β_5	۰/۰۲۹	۰/۰۳۵	۰/۴۰۴
β_6	۱/۰۵۶	۰/۶۷۱	۰/۱۲۴
σ^2	۲۰۳/۶۳۲	۱/۷۶۶	۰/۰۰۰۵



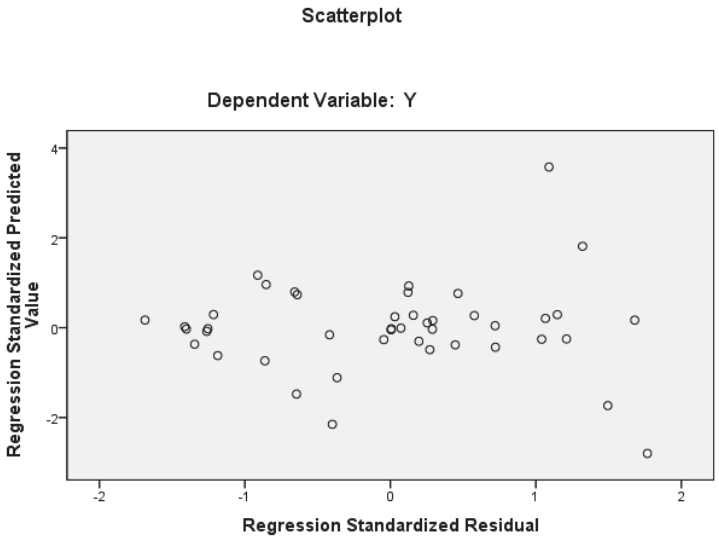
شکل ۳: نمودار بافت‌نگار باقی مانده‌های مدل (۱۳)

باتوجه به شکل‌های فوق می‌توان نتیجه گرفت که مدل نرمال برای خطاها در این مثال مناسب نیست.

نکته: آزمون شاپیرو- ویلک برای داده‌هایی جهت تست نرمالیتی با حجم کم، به عنوان مثال ۵۰ و یا کمتر، به کار می رود و آزمون کلموگروف-اسمیرنوف برای داده‌هایی با حجم نمونه بالا استفاده می شود. برای بررسی نرمال بودن مانده‌ها میزان $p-value$ متناظر با آن‌ها را بررسی می‌کنیم. چنانچه میزان $p-value$ این آزمون ها بیشتر از ۰/۰۵ باشد نتیجه می‌گیریم که مانده‌ها از توزیع نرمال تبعیت می کنند.

در ادامه آزمون نرمال بودن را برای مانده‌های مدل (۱۳) انجام دادیم. نتایج حاصل در جدول ۲ نشان داده شده اند که نیز گواه بر صحت ادعای فوق می‌باشد. زیرا با توجه به میزان $p-value$ در جدول ۲، فرض نرمال بودن مانده‌های مدل در سطح ۰/۰۵ رد می‌شود.

همچنین شکل ۵ نمودار چندک-چندک نرمال برای باقیمانده ها را نشان می دهد .



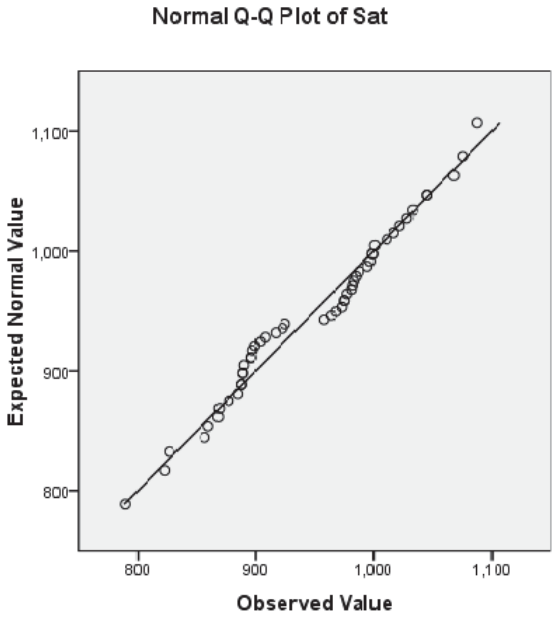
شکل ۴: نمودار مانده ها در برابر مقادیر برازش شده

جدول ۲: آزمون شاپیرو - ویلک برای نرمال بودن ماندهها

$p-value$	درجه آزادی	آماره
۰/۰۲۳	۴۴	۰/۹۴

حال فرض می‌کنیم مؤلفه‌های خطا دارای توزیع اسلش هستند و برآورد حداکثر درستنمایی پارامترهای مدل را به دست می آوریم و مشاهده می‌کنیم که مدل اسلش برازش بهتری برای این مدل رگرسیونی است. در مسئله برآوردیابی حداکثر درستنمایی پارامترهای آن از نرم افزار R استفاده شده است که نتایج در جدول ۳ ارائه شده است.

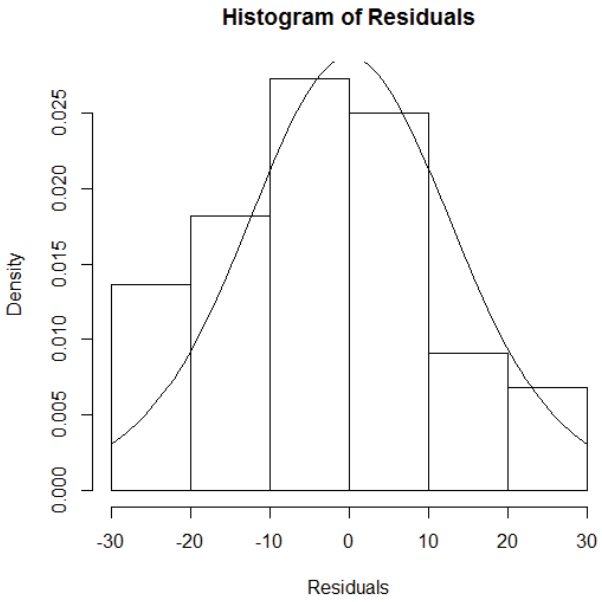
نمودار بافت نگار باقی ماندههای مدل با فرض برقراری توزیع اسلش در شکل ۶ رسم شده است.



شکل ۵: نمودار چندک-چندک نرمال برای ماندههای مدل (۱۳)

جدول ۳: برآورد پارامترهای مدل (۱۳) با فرض برقراری توزیع اسلش برای خطاهای مدل

پارامترها	برآورد پارامترها	انحراف معیار	$loglike$
β_0	$-۶/۱۳۳$	۱۰/۱۶۶	$-۱۷۵/۵۷۷$
β_1	۰/۳۲۹	۰/۲۰۵	
β_2	۰/۹۸۵	۰/۳۰۶	
β_3	$-۰/۲۷۹$	۰/۱۱۰	
β_4	۰/۶۵۷	۰/۳۳۱	
β_5	۰/۰۱۱	۰/۰۲۸	
β_6	$-۰/۴۲۳$	۰/۵۲۲	
β_7	۱۰۸/۷۸۴	۱/۸۴۶	
q	۳/۰۱۰	۰/۹۳۳	



شکل ۶: نمودار بافت نگار باقی مانده‌های مدل (۱۳) با فرض برقراری توزیع اسلش برای خطاها

برآورد واریانس (σ^2) در مدل نرمال از مدل اسلش بیشتر است در واقع ما انتظار چنین مسئله ای را داشتیم چون پراکندگی زیاد تا حدودی جای خود را به دم سنگینی می‌دهد و به خاطر اینکه توزیع اسلش دم سنگین تر از توزیع نرمال است، این امر کاملا معقول به نظر می‌رسد.

اکنون با معیارهای کمی به مقایسه دو مدل می‌پردازیم.

تعداد زیادی از معیارهای انتخاب مدل بر اساس نظریه ی اطلاع پدید آمده‌اند. یکی از معروفترین معیارها لگاریتم درستنمایی است، هنگامی‌که در جستجوی بهترین برازش برای داده‌ها در بین سایر توزیع‌ها هستیم معمولاً توزیع با بیشترین مقدار لگاریتم درستنمایی به عنوان برازش بهتر انتخاب می‌شود. زمانی که تعداد پارامتر های برآورد شده یکسان نباشد، این معیار چندان مناسب نیست. به طور معمول برای مقایسه مدل‌هایی با تعداد پارامتر متفاوت، از معیار اطلاع آکائیک (AIC) و معیار اطلاع بیزي (BIC) مناسب هستند که به ترتیب توسط آکائیک [۱۷] [۱] و شوارتز [۱۸] [۱۷] برای انتخاب بهترین مدل آماری معرفی شدند. مقدار معیار اطلاع آکائیک (AIC) با استفاده از فرمول $AIC = -۲\ell(\widehat{\theta}) + ۲p^*$ که در آن p^* تعداد پارامترهای مدل است، معین می‌شود. معیار اطلاع بیزي (BIC) به صورت زیر تعریف می‌شود:

$BIC = -۲\ell(\widehat{\theta}) + p^* lnn$ که در آن n حجم نمونه است. هنگامی‌که در جستجوی برازش بهتر برای داده‌ها در بین سایر توزیع‌ها هستیم توزیع با کوچکترین مقدار AIC و BIC به عنوان برازش بهتر انتخاب می‌شود. معیارهای

[۱۷] Akaike Hirotugu

[۱۸] Schwarz

لگاریتم درستنمایی، اطلاع آکائیک و اطلاع بیزی برای دو حالت مدل با خطای توزیع نرمال و با خطای توزیع اسلش در جدول ۴ ارائه شده است.

جدول ۴: مقادیر معیارهای لگاریتم درستنمایی، اطلاع آکائیک و اطلاع بیزی برای دو مدل با خطاهای توزیع اسلش و توزیع نرمال

معیار سنجش مدل	مدل نرمال	مدل اسلش
AIC	۳۷۸/۶۱۹	۳۶۷/۱۵۴
BIC	۳۹۰/۴۸۶	۳۸۱/۴۲۷۵
لگاریتم درستنمایی	−۱۸۰/۳۰۹	−۱۷۵/۵۷۷

از بین دو مدل اسلش و نرمال، مدل با خطای توزیع اسلش دارای کمترین مقدار معیار اطلاع آکائیک (AIC) و معیار اطلاع بیزی (BIC) و بیشترین مقدار لگاریتم درستنمایی است، پس نسبت به توزیع نرمال برازش بهتری برای خطاهای مدل رگرسیونی در رابطه (۱۳) می باشد.

۷ نتیجه گیری

در این مقاله توزیع اسلش و برخی ویژگی های آن را مرور نمودیم و از آن در برازش مدل‌های رگرسیونی چندگانه معمولی استفاده کردیم و با استفاده از داده های واقعی نشان دادیم که توزیع اسلش جایگزین مناسبی برای توزیع نرمال در خطای مدل های رگرسیونی است. برای مقایسه دو مدل از معیارهای مختلف استفاده شد و مشاهده کردیم که توزیع اسلش نسبت به توزیع نرمال برای مؤلفه‌های خطای مدل رگرسیونی (۱۳) مناسب‌تر است.

مراجع

[1] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723.

[2] Andrews, D.F. and Mallows, C.L. (1974). Scale Mixtures of Normal Distri- butions. Journal of the Royal Statistical Society, B36, 99-102.

[3] Arslan, O. and Genc, A.I. (2009). A generalization of the multivariate slash distribution. Journal of Statistical Planning and Inference, 139, 1164-1170.

[4] Fisher, R.A. (1956). Statistical Methods in Scientific Inference. Oliver and Boyed, London.

[5] Genç, A.I. (2007). A generalization of the univariate slash by a scale-mixtured exponential power distribution. Journal of Communications in Statistics-Simulation and Computation, 36, 937-947.

[6] Genç, A.I . (2012). A skew extension of the slash distribution via beta-normal distribution. Statist. Papers, 54(2), 427-442.

[7] Gomez, H.W., Quintana, A. and Torres, J. (2007). A new family of slash-distributions with elliptical contours. Statistics and Probability Letters, 77, 717-725.

[8] Gómez, W., Reyesa, H. and Bolfarineb, H. (2013). Modified slash distribution. Statistics: A Journal of Theoretical and Applied Statistics, 47(5), 929-941.

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

[10] Johnson, N.L., Kotz, S. and Balakrishan, N. (1995). Continuouse univariate distribution, Vol. 1, Second ed., Wiley, New York.

[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] Lange, K. and Sinsheimer, J.S. (1993). Normal Independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics, 2, 175-198.

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

[15] Rencher, A.C. and Schaalje, G. B. (2008). Linear models in statistics. 2nd ed., Wiley, New York.

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

[17] Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6(2), 461–464.