# مدل‌های رگرسیونی برای تحلیل داده‌های چوله دو مدی

**حسن میرزاوند [۱]، مجید جعفری خالدی [۲]**

*تاریخ دریافت: ۱۴۰۰/۱۰/۲۰*
*تاریخ پذیرش: ۱۴۰۱/۰۵/۲۴*

چکیده:

برای استنباط آماری در مورد پارامترهای مدل رگرسیونی نیاز به فرض توزیع مشخصی بر روی عبارت خطای تصادفی می‌باشد. یک فرض اساسی در مدل رگرسیون خطی این است که عبارت خطای تصادفی از یک توزیع نرمال پیروی کند. بااین‌حال، در پژوهش‌های آماری گاهی با داده‌هایی مواجه می‌شویم که توزیع آن‌ها چولگی و دو مدی را ارائه می‌دهند، و دیگر نمی‌توان از فرض توزیع نرمال برای تحلیل آن‌ها استفاده کرد. یک رویکرد مرسوم برای حل این مسئله به‌کارگیری آمیخته‌ای از مدل‌های چوله نرمال است. اما در این‌گونه مدل‌ها تعداد پارامترها به نحو فزاینده‌ای افزایش می‌یابد که این خود برازش مدل‌ها به داده‌ها را دشوار می‌نماید. بعلاوه مدل‌های آمیخته خود درگیر مسائلی مانند شناساناپذیری هستند. در این حالت یک‌راه‌حل مناسب استفاده از توزیع‌های منعطفی است، که بتوانند چولگی و دو مدی بودن داده‌ها را در مدل بندی لحاظ کنند. تاکنون روش‌های مختلفی ارائه‌شده که بر مبنای توسعه توزیع چوله‌نرمال، توزیع‌های دو مدی نامتقارن ایجادشده‌اند. در این مقاله ازاین‌روش‌ها برای ساخت و معرفی مدل رگرسیونی منعطف نسبت به مدل‌های رگرسیون مبتنی بر توزیع چوله‌نرمال و آمیخته‌ای از دو توزیع چوله‌نرمال استفاده‌شده و با به‌کارگیری مثال شبیه‌سازی عملکرد آن‌ها موردبررسی قرار می‌گیرد. سپس نحوه کاربست آن‌ها در یک مثال کاربردی مربوط به مجموعه داده‌های اسب‌دوانی نشان داده می‌شود.

**واژه‌های کلیدی:** چولگی، توزیع‌های دو مدی، تقارن، توزیع‌های آمیخته، رگرسیون.

## ۱ مقدمه

نظریه سنتی تحلیل رگرسیون بر پایه نرمال بودن توزیع مشاهدات در سطوح مختلف متغیر کمکی بناشده است، در این صورت فرض می‌شود که عبارت خطای تصادفی مدل از توزیع نرمال پیروی می‌کند.

**تعریف ۱.۱.** *یک مدل رگرسیون خطی با باقیمانده‌های نرمال که با نماد $N-LR$ نمایش داده می‌شود، به‌صورت*

$$Y_i = x_i^\top \beta + \epsilon_i, \qquad i = 1, \cdots, n \qquad (1)$$

*تعریف می‌شود، که در آن $Y_i$ متغیر پاسخ، $x_i = (1, x_{i1}, \cdots, x_{ip})^\top$ بردار مقادیر متغیر مستقل با بُعد $(p+1)\times 1$، $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ بردار ضرایب رگرسیونی و خطای تصادفی $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. در این صورت با توجه به خواص توزیع نرمال، متغیر پاسخ $Y_i|x_i \overset{iid}{\sim} N(x_i^\top \beta, \sigma^2)$ که تابع چگالی آن به‌صورت*

$$f_{Y_i}(y_i|x_i, \theta) = \phi\left(\frac{y_i - x_i^\top \beta}{\sigma}\right); \qquad y \in \mathbb{R} \qquad (2)$$

*که در آن $\phi$ تابع چگالی استاندارد و $\theta = (\beta^\top, \sigma)$ بردار پارامترهای مدل است.*

این در حالی است که در بسیاری از کاربردها مشاهدات انحراف زیادی از توزیع نرمال دارند و مدل‌سازی بر پایه فرض نرمال به برآوردهای غیرمنطقی از پارامترهای مدل منجر می‌شود. این خصوصیات را می‌توان با تبدیل داده‌ها مرتفع نمود، که نرمال بودن تقریبی مشاهدات را نتیجه می‌دهد. بااین‌حال برخی از مشکلات این روش به شرح زیر هستند:

- پارامترها ممکن است تفسیرپذیری را در مقیاس تبدیل‌شده از

[۱] گروه آمار، دانشکده علوم ریاضی، دانشگاه تربیت مدرس
[۲] عضو هیئت‌علمی گروه آمار، دانشکده علوم ریاضی، دانشگاه تربیت مدرس (نویسنده مسئول: jafari-m@modares.ac.ir)

دست بدهند.

- گاهی وقت‌ها پیدا کردن تبدیل مناسب زمان‌بر است و بعضاً ممکن است تبدیلی که داده‌ها را نرمال نماید، یافت نشود.

ازاین‌رو از دیدگاه عملی، نیاز به ارائه یک مدل نظری مناسب است، که از تبدیل داده‌ها جلوگیری کند. جایگزین کردن توزیع نرمال با توزیع‌های مناسب‌تری که قابلیت مدل بندی داده‌های نامتقارن را نیز داشته باشند، راهکاری است که در سال‌های اخیر مورد توجه بسیاری از محققین قرارگرفته است. در این راستا، آزالینی [۴] یک نتیجه اساسی را برای توسعه مدل‌های چوله-متقارن ارائه می‌کند.

**لم ۲.۱.** *فرض کنید $f_0$ یک تابع چگالی متقارن حول صفر باشد و $G$ یک تابع توزیع به‌طوری‌که $G'$ تابع چگالی آن متقارن حول صفر است. در این صورت:*

$$f_Z(z \mid \alpha) = 2 f_0(z) G(\alpha z) \quad -\infty < z < \infty \qquad (3)$$

برای هر $\alpha \in \mathbb{R}$ یک تابع چگالی است. از نماد $Z \sim Sf_0(\alpha)$ برای نشان دادن این واقعیت که $Z$ دارای چگالی (۳) است، استفاده خواهد شد.

با استفاده از لم ۲.۱ می‌توان طیف گسترده‌ای از توزیع‌های نامتقارن تولید کرد. که از جمله مشهورترین آن‌ها توزیع چوله‌نرمال است.

**تعریف ۳.۱.** *اگر $Z \sim SN(\alpha)$ باشد، تابع چگالی آن به‌صورت*

$$f(z \mid \alpha) = 2\phi(z)\Phi(\alpha z) \quad z \in \mathbb{R} \qquad (4)$$

می‌باشد، که در آن $\phi$ و $\Phi$ به ترتیب تابع چگالی و تابع توزیع نرمال استاندارد می‌باشند. $\alpha$ پارامتر کنترل چولگی است، به ازای $\alpha = 0$ توزیع متقارن و به چگالی نرمال استاندارد تبدیل می‌شود. لازم به ذکر است تابع چگالی (۴) به ازای مقادیر مختلف پارامتر $\alpha$ تغییر می‌کند، که نشان‌دهنده شناساپذیر بودن این خانواده از توزیع‌ها است. به بیان دقیق‌تر، به ازای $\alpha_1, \alpha_2 \in \mathbb{R}$ به‌طوری‌که $\alpha_1 \neq \alpha_2$ آنگاه $f(z \mid \alpha_1) \neq f(z \mid \alpha_2)$ می‌باشد. توزیع چوله‌نرمال یک متغیره نخستین بار توسط آزالینی [۴] ارائه شد. هنز [۱۰] یک روش ساخت متغیرهای تصادفی با توزیع چوله‌نرمال معرفی کرد، سپس گشتاورهای آن را به دست آورد. این توزیع علی‌رغم ویژگی‌های خوبی که دارد، در تخمین پارامتر شکل $\alpha$ با مشکلاتی مواجه است. به‌طور خاص، برای حجم‌های نمونه متوسط، برآوردگر ماکسیمم درستنمایی با احتمال مثبت، نامتناهی است ([۱۵] را مشاهده کنید). [۱۶] از یک تابع امتیاز اصلاح‌شده به‌عنوان معادله درستنمایی برای برآورد پارامتر شکل استفاده می‌کند و نشان می‌دهد که برآوردگر ماکسیمم درستنمایی اصلاح‌شده همیشه متناهی است.

**تعریف ۴.۱.** *اگر $Z \sim SN(\alpha)$ باشد، آنگاه تابع مولد گشتاور آن به‌صورت*

$$M_Z(t) = 2\exp\{\frac{t^2}{2}\}\Phi(\frac{\alpha}{\sqrt{1+\alpha^2}}t), \qquad (5)$$

*است.*

با استفاده از تابع مولد گشتاور به‌سادگی گشتاورهای اول و دوم به‌صورت

$$E(Z) = \frac{\partial}{\partial t} M_Z(t)|_{t=0} \Rightarrow E(Z) = \sqrt{\frac{2}{\pi}}\lambda,$$
$$E(Z^2) = \frac{\partial}{\partial t^2} M_Z(t)|_{t=0} = 1 \Rightarrow Var(Z) = 1 - \frac{2}{\pi}\lambda^2,$$

به دست می‌آیند، که در آن $\lambda = \frac{\alpha}{\sqrt{1+\alpha}}$ است. $E(Z)$ تابعی صعودی از $\lambda$ است و $Var(Z)$ تابعی نزولی از $|\lambda|$ است. آزالینی [۴] ضرایب چولگی و کشیدگی متغیر تصادفی چوله‌نرمال را به ترتیب به‌صورت

$$\sqrt{\beta_1} = \frac{1}{2}(4-\pi)(\frac{E^2(Z)}{Var(Z)})^{\frac{3}{4}},$$
$$\beta_2 = 2(\pi-3)(\frac{E^2(Z)}{Var(Z)})^2,$$

محاسبه نمود. بر این اساس آزالینی [۴] نشان داد که مقدار ضریب چولگی و ضریب کشیدگی در بازهٔ $(-0/995, 0/995)$ و $(-0/869, 0/869)$ تغییر می‌کند. اضافه کردن پارامترهای مکان و مقیاس به توزیع چوله‌نرمال، انعطاف‌پذیری آن را بیشتر می‌کند و با تغییر مقدار پارامترها می‌توان کنترل بیشتری روی توزیع داشت. تبدیل $X = \mu + \sigma Z$، با $\mu \in R$ و $\sigma > 0$ را در نظر می‌گیریم. متغیر تصادفی $X$ دارای توزیع چوله‌نرمال با تابع چگالی

$$f_X(x) = \frac{2}{\sigma}\phi(\frac{x-\mu}{\sigma})\Phi(\alpha(\frac{x-\mu}{\sigma})), \qquad (6)$$

در ادامه مدل رگرسیون مبتنی بر این توزیع $SN$ معرفی می‌شود.

**تعریف ۵.۱.** *یک مدل رگرسیون خطی با باقیمانده‌های چوله‌نرمال که با نماد $SN-LR$ نمایش داده می‌شود، به‌صورت*

$$Y_i = x_i^\top \beta + \epsilon_i, \qquad i = 1, \cdots, n \qquad (7)$$

*تعریف می‌شود، که در آن $Y_i$ متغیر پاسخ، $x_i = (1, x_{i1}, \cdots, x_{ip})^\top$ بردار مقادیر متغیر مستقل با بعد $1 \times (p+1)$، $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ بردار ضرایب رگرسیونی و خطای تصادفی $\epsilon_i \overset{iid}{\sim} SN(0, \sigma, \alpha)$. در*

*این صورت با توجه به خواص توزیع $SN$، متغیر پاسخ $Y_i|x_i \overset{iid}{\sim} SN(x_i^\top\beta,\sigma,\alpha)$ که تابع چگالی آن به‌صورت*

$$f_{Y_i}(y_i|x_i,\theta)=\frac{2}{\sigma}\phi\left(\frac{y_i-x_i^\top\beta}{\sigma}\right)\Phi\left(\alpha\frac{y_i-x_i^\top\beta}{\sigma}\right);\qquad y\in\mathbb{R} \tag{8}$$

*که در آن $\phi$ تابع چگالی استاندارد و $\theta=(\beta^\top,\sigma,\alpha)$ بردار پارامترهای مدل است. امید ریاضی و واریانس $Y_i$ به‌صورت*

$$E(Y_i|x_i)=x_i^\top\beta+c\alpha$$

$$Var(Y_i|x_i)=\sigma+(1-c^2)\alpha^2$$

*که در آن $c=\sqrt{2/\pi}$ می‌باشد.*

این مدل رگرسیونی توسط افراد مختلفی موردمطالعه قرارگرفته است، که از جمله می‌توان به مطالعات آرلانو-واله و همکاران [۳]، لاچوس و همکاران [۱۱] و کانچو و همکاران [۶] اشاره کرد، که از توزیع چوله‌نرمال برای برازش مدل رگرسیونی استفاده کرده‌اند. اهمیت توزیع چوله‌نرمال و مبنا قرارگرفتن آن در بسیاری از مباحث استنباط آماری و مدل‌سازی به این دلیل است که این توزیع علی‌رغم داشتن قابلیت مدل بندی مشاهدات نامتقارن، شباهت زیادی به توزیع نرمال داشته و آن را نه به‌عنوان یک توزیع حدی، بلکه به‌عنوان یک عضو در بردارد. توجه کنید که مدل رگرسیون مبتنی بر توزیع چوله‌نرمال تک مدی است، اما در عمل داده‌هایی وجود دارند که نیازمند برازش با توزیع‌های نامتقارن دومدی هستند، در این صورت این مدل ازکارایی لازم برخوردار نیست. به‌عنوان‌مثال داده‌های حداقل دمای کشور در بهمن‌ماه سال ۱۳۹۶ جمع‌آوری‌شده در ۳۸۱ ایستگاه هواشناسی را در نظر بگیرید. شکل ۱ موقعیت جغرافیایی ایستگاه‌ها و هیستوگرام حداقل دمای هوا را به نمایش می‌گذارد. هیستوگرام نشان می‌دهد توزیع داده‌ها چوله به راست بوده و دو مدی است. لازم به ذکر است که می‌توان از متغیر ارتفاع ایستگاه‌ها از سطح دریا به‌عنوان متغیر کمکی در تحلیل رگرسیونی این داده‌ها استفاده نمود. اکنون سؤالی که مطرح می‌شود این است که چگونه می‌توان چولگی و دومدی موجود در داده‌ها را مدل بندی نمود؟ ادامه این مقاله به‌صورت زیر سازمان‌دهی شده است. در بخش ۲، مدل رگرسیون مبتنی برآمیخته‌ای از دو توزیع چوله نرمال مطالعه می‌گردد. در ادامه در بخش ۳، توزیع‌های نامتقارن دو مدی معرفی می‌گردد، سپس برخی از خواص و ویژگی‌های آن‌ها موردمطالعه قرار می‌گیرند. در بخش ۴، مدل رگرسیون مبتنی بر توسعه‌های دو مدی توزیع چوله‌نرمال معرفی می‌شود. در بخش ۵، نحوه کاربست مدل‌های رگرسیونی توسعه‌یافته، با استفاده از مطالعات شبیه‌سازی و همچنین تحلیل مجموعه داده‌های اسب‌دوانی ($horses$) نشان داده می‌شود. همچنین در بخش ۶، به بحث و نتیجه‌گیری پرداخته خواهد شد.

## ۲ مدل رگرسیونی مبتنی برآمیخته‌ای از دو توزیع چوله‌نرمال

یکی از راهکارهایی که به‌منظور مدل‌سازی چولگی و دو مدی می‌توان استفاده نمود به‌کارگیری مدل‌های آمیخته متناهی است، که در ادامه به معرفی آن‌ها پرداخته‌شده است. توزیع‌های آمیخته متناهی، یعنی توزیع‌هایی که به‌صورت مجموع موزون از چند توزیع با مؤلفه‌های ساده‌تر نوشته می‌شود، بسیار مورد توجه قرار دارند. معمولاً توزیع مؤلفه‌ها با توجه به نوع داده‌های موردبررسی و نیز به دلیل جذابیت‌های محاسباتی به کلاس‌های خاصی از تابع چگالی‌های پارامتری محدود می‌شوند. به‌طورکلی مدل آمیخته متناهی بر اساس آمیخته موزونی از تابع چگالی‌های دلخواه به‌صورت

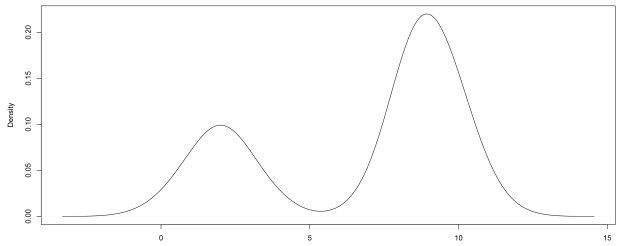
$$f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\eta})=\sum_{k=1}^{K}\eta_k f_k(\boldsymbol{y}|\boldsymbol{\theta}_k), \tag{9}$$

تعریف می‌شود که در آن $\boldsymbol{y}=(y_1,\ldots,y_n)'$ بردار مشاهدات، $K$ تعداد مؤلفه‌ها و $\eta_k$ها وزن‌های آمیزنده[3] مؤلفه‌ها هستند، به‌طوری‌که

$$\forall k:\ \ \eta_k\geq 0,\ \ \sum_k\eta_k=1.$$

در رابطه (۹)، $\boldsymbol{\eta}=(\eta_1,\ldots,\eta_K)'$ بردار وزن‌های آمیزنده و $\boldsymbol{\theta}=(\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_K')'$ بردار پارامترهای نامعلوم تابع چگالی‌ها را نمایش می‌دهند. همچنین $f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)$ تابع چگالی مؤلفه $k$ام با پارامترهای $\boldsymbol{\theta}_k$ است. در رابطه (۹) تعداد مؤلفه‌ها ثابت در نظر گرفته‌شده است، اما در عمل تعداد مؤلفه‌ها نامعلوم است و بایستی به همراه وزن‌های آمیزنده و پارامترهای موجود در تابع چگالی مؤلفه‌ها، از روی‌داده‌های موجود تعیین و برآورد شوند. برای درک بهتر از مدل آمیخته متناهی، آمیخته‌ای از توزیع‌های نرمال در شکل ۲ نمایش داده‌شده است.

[3] Mixing Weights




**شکل ۲.** نمودار تابع چگالی آمیخته از دو توزیع نرمال با وزن‌های ۰/۳ و ۰/۷

یکی از پرکاربردترین مدل‌های آمیخته، برای زمانی که گروه‌های متفاوت از داده‌ها دارای چولگی باشند، مدل آمیخته‌ای از چوله‌نرمال‌ها است که جایگزین مناسب‌تری نسبت به مدل آمیخته‌ای از نرمال‌ها است. آرلانو-واله و همکاران [۲] یک مدل چوله-متقارن معرفی می‌کنند، که آمیخته‌ای از توزیع‌های چوله‌نرمال است.

**تعریف ۱.۲.** *مدل آمیخته متناهی چوله‌نرمال دو مؤلفه‌ای، آمیخته‌ای از ۲ توزیع چوله‌نرمال یک متغیره به‌صورت*

(۱۰)
$$f(y|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\eta)=\eta f_{SN}(y;\mu_1,\sigma_1^2,\alpha_1)+(1-\eta)f_{SN}(y;\mu_2,\sigma_2^2,\alpha_2)$$

*است که در آن*

$$f_{SN}(y;\mu_k,\sigma_k^2,\alpha_k)=\frac{2}{\sigma_k}\phi(\frac{y-\mu_k}{\sigma_k})\Phi(\alpha_k(\frac{y-\mu_k}{\sigma_k}))\qquad k=1,2$$

*تابع چگالی گروه $k$ام با پارامترهای $\boldsymbol{\theta}_k=(\mu_k,\sigma_k,\alpha_k)$ که در آن $\mu_k$ پارامتر مکان، $\sigma_k$ پارامتر مقیاس و $\alpha_k$ پارامتر چولگی مربوط به گروه $k$ام هستند، این توزیع با نماد $Y\sim Mix.SN(\mu_1,\mu_2,\sigma_1,\sigma_2,\alpha_1,\alpha_2,\eta)$ نمایش داده می‌شود. برای مطالعه بیشتر پیرامون آمیخته‌ای از توزیع‌های چوله‌نرمال و چوله‌تی به لین و همکاران [۱۲] و بهرامی [۱] مراجعه کنید.*

از طرفی، هنگامی‌که جامعه ساختاری ناهمگن دارد و از زیرجوامعی چوله تشکیل‌شده است،(به‌عنوان‌مثال، داده‌های درآمد خانوارهای روستایی ایران تحلیل‌شده توسط بهرامی [۱])، می‌توان آمیخته‌ای متناهی از توزیع‌های چوله‌نرمال را به‌منظور مدل‌سازی خطای تصادفی استفاده نمود. در این حالت، می‌توان از مدل رگرسیونی مبتنی برآمیخته‌ای از چوله‌نرمال‌ها برای خطای تصادفی به‌عنوان جایگزین مدل سنتی استفاده کرد.

**تعریف ۲.۲.** *فرض کنید، خطای تصادفی مدل دارای توزیع آمیخته‌ای از دو توزیع چوله‌نرمال باشد که با نماد $\epsilon_i \overset{iid}{\sim} Mix.SN(0,0,\sigma_1,\sigma_2,\alpha_1,\alpha_2,\eta_1,\eta_2)$ نمایش داده می‌شود، در این صورت متغیر پاسخ دارای توزیع آمیخته دو مؤلفه‌ای از توزیع چوله‌نرمال به‌صورت*

$$Y_i|x_i\sim\sum_{j=1}^{2}\eta_j f_{SN}(y_i|\theta_j(x_i))\qquad i=1\cdots n\qquad(11)$$

*است، که با نماد $Mix.SN-LR$ نمایش داده می‌شود. در رابطه (۱۱) بردار $\theta_j(x_i)=(\mu_j(x_i),\sigma_j^2,\alpha_j)$، $x_i=\mu_j(x_i)=x_i^\top\beta_j$ بردار $(1,x_{i1},\cdots,x_{ip})^\top$ و $\beta_j=(\beta_{j0},\beta_{j1},\cdots,\beta_{jp})^\top$ بردار ضرایب رگرسیونی زیر جامعه $j$ام و $\eta=(\eta_1,\eta_2)$ بردار ضرایب آمیختگی را نشان می‌دهد.*

اما همان‌طور که بیان شد به دلیل بُعد زیاد پارامترها و مشکلات شناساناپذیری مبتلابه این‌گونه مدل‌ها استفاده از آن‌ها چالش‌برانگیز است. مک‌لاچان و پیل [۱۴] و مارین و همکاران [۱۳] مشکلات جدی را در مورد برآورد چنین مدل‌هایی ارائه می‌دهند. حال سؤالی که در اینجا مطرح می‌شود این است که چگونه مدل رگرسیون منعطفی معرفی کنیم که هم چولگی و دو مدی بودن را کنترل کند و هم از مشکلات مبتلابه مدل‌های آمیخته رنج نبرد؟

یک‌راه حل دیگر برای تحلیل رگرسیونی زمانی که چولگی و دو مدی به‌طور هم‌زمان وجود دارد، استفاده از توسعه‌های دو مدی توزیع چوله‌نرمال است که در بخش ۳ معرفی می‌شوند.

# ۳ معرفی توزیع‌های نامتقارن دو مدی

در این بخش، توسعه‌هایی از توزیع چوله‌نرمال به حالت دو مدی، بررسی می‌شود. از جمله خانواده جدیدی از توزیع‌های چوله‌نرمال انعطاف‌پذیر[4] بانام آلفا ـ چوله‌نرمال که توسط اِلال-اُلیویرو [۷] معرفی شد، که داده‌های تک مدی و دو مدی را پوشش می‌دهد. به‌صورت زیر معرفی کردند.

**تعریف ۱.۳.** *اگر متغیر تصادفی $Z$ دارای تابع چگالی*

$$f(z;\alpha)=\frac{1+(1-\alpha z)^2}{2+\alpha^2}\phi(z)\qquad z\in\mathbb{R}\qquad(12)$$

*که در آن $\alpha\in\mathbb{R}$، در این صورت $Z$ را متغیر تصادفی آلفا-چوله‌نرمال با پارامتر $\alpha$ نامیده و به‌صورت $Z\sim ASN(\alpha)$ نشان می‌دهند. در مدل $ASN(\alpha)$ چون تنها از یک پارامتر برای کنترل چولگی و مد توزیع استفاده می‌شود، چندان کارآمد به نظر نمی‌رسد.*

[4] Skew-Flexible-Normal Distributions

رده توزیع‌های معرفی‌شده توسط اِلال-اُلیویرو [۷]، منجر به تعمیم و بسط سایر توزیع‌ها توسط محققان دیگری شد، که در ادامه به آن‌ها پرداخته می‌شود. در ساخت و معرفی توسعه‌های دو مدی توزیع چوله‌نرمال، فرع ۲.۳ نقش اساسی ایفا می‌کند.

**نتیجه ۲.۳.** *( آزالینی و کاپتانو [۵]) با فرض آنکه $f_0$، $G'$ و $W$ چنان باشند که در لم ۲.۱ بیان شد و $X \sim G'$ ، $Y \sim f_0$ متغیرهای تصادفی مستقل باشند، آنگاه متغیر تصادفی $Z$، که به‌صورت*

$$Z = \begin{cases} Y & X \le W(Y) \\ -Y & X > W(Y) \end{cases}$$

*تعریف می‌شود، دارای تابع چگالی (۳) است.*

- اگر در فرع ۲.۳، قرار دهیم W(Y)=αY، که در آن $\alpha$ یک عدد حقیقی است و $X \sim N(0,1)$ و همچنین $Y$ دارای توزیع نرمال بریده‌شده $Y \sim c_\delta \phi(y+\delta) I(y \ge 0)$ باشد، در این صورت متغیر تصادفی $Z$ دارای توزیع چوله‌نرمال دو مدی معرفی‌شده توسط گومز و همکاران [۹] است.
- اگر در فرع ۲.۳، قرار دهیم W(Y)=αY، که در آن $\alpha$ یک عدد حقیقی است و $X \sim N(0,1)$ و $Y$ دارای توزیع نرمال دو مدی $Y \sim BN(\delta)$ باشد، در این صورت متغیر تصادفی $Z$ دارای توزیع چوله‌نرمال دو مدی معرفی‌شده توسط اِلال-اُلیویرو و همکاران [۸] است.

در ادامه به‌تفصیل این دو توسعه توزیع چوله‌نرمال موردبحث و بررسی قرار می‌گیرد.

## ۱.۳ توزیع چوله‌نرمال انعطاف‌پذیر $(SFN)$

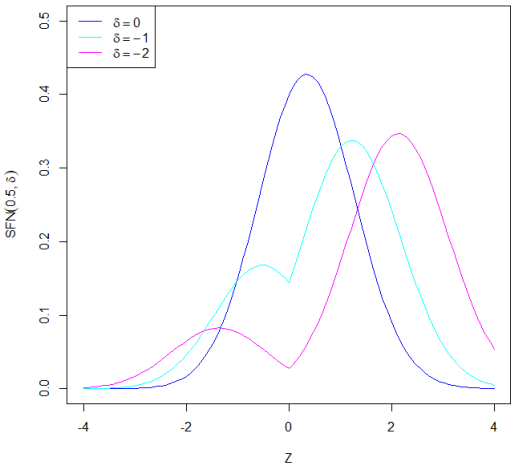
**تعریف ۳.۳.** *(گومز و همکاران [۹]) متغیر تصادفی $Z$ دارای توزیع چوله‌نرمال-انعطاف‌پذیر با پارامترهای $\alpha$ و $\delta$ نامیده می‌شود، هرگاه تابع چگالی آن به‌صورت*

$$f(z|\alpha,\delta) = c_\delta \phi(|z|+\delta)\Phi(\alpha z); \qquad z \in \mathbb{R} \qquad (۱۳)$$

*یا*

$$f(z|\alpha,\delta) = \begin{cases} c_\delta \phi(z-\delta)\Phi(\alpha z) & z \le 0 \\ c_\delta \phi(z+\delta)\Phi(\alpha z) & z > 0 \end{cases} \qquad (۱۴)$$

باشد، که در آن $\phi$ و $\Phi$ به ترتیب نشان‌دهنده تابع چگالی و تابع توزیع نرمال استاندارد، $\alpha, \delta \in \mathbb{R}$ و $c_\delta^{-1} = (1 - \Phi(\delta))$ است. این توزیع با نماد $Z \sim SFN(\alpha,\delta)$ نمایش داده می‌شود. لازم به ذکر است تابع چگالی (۱۴) به ازای مقادیر مختلف پارامترهای $\alpha$ و $\delta$ تغییر می‌کند، که نشان‌دهنده شناساپذیر بودن این توزیع است (گومز و همکاران [۹]).




**شکل ۱.۳.** نمایش توزیع $SFN(\alpha,\delta)$ به ازای مقادیر مختلف پارامترهای $\alpha$ و $\delta$.

شکل ۱.۳ نمایش گرافیکی توزیع $SFN(\alpha,\delta)$، برای مقادیر مختلف پارامترها را نشان می‌دهد. با توجه به این شکل، پارامتر $\alpha$ پارامتر چولگی و $\delta$ پارامتر دو مدی ساز به نحوی است که به ازای مقادیر کمتر از صفر دو مدی بودن را برای توزیع به ارمغان می‌آورد. خواص زیر مستقیماً از تعریف ۳.۳ به دست می‌آیند.

**گزارهٔ ۴.۳.** *فرض کنید متغیر تصادفی $Z$ دارای توزیع $SFN(\alpha,\delta)$ باشد، در این صورت*

- *اگر $\alpha = 0$ و $\delta = 0$ آنگاه $Z \sim N(0,1)$*
- *اگر $\delta = 0$ آنگاه $Z \sim SN(\alpha)$*
- *اگر $\alpha = 0$ آنگاه $f(z|\alpha=0,\delta) = c_\delta \phi(|z|+\delta)/2$*
- *اگر $\alpha \to +\infty$ آنگاه $f(z|\alpha,\delta) = c_\delta \phi(z+\delta) I(z \ge 0)$*
- *اگر $\alpha \to -\infty$ آنگاه $f(z|\alpha,\delta) = c_\delta \phi(z-\delta) I(z < 0)$*

**گزارهٔ ۵.۳.** *فرض کنید $Z \sim SFN(\alpha,\delta)$ است. اگر $\delta > 0$، در این صورت $Z$ یک متغیر تصادفی با تابع چگالی دو مدی می‌باشد.*

**نتیجه ۶.۳.** *فرض کنید $Z \sim SFN(\alpha,\delta)$ است. اگر $\alpha = 0$ و $\delta > 0$، آنگاه $Z$ یک متغیر تصادفی با تابع چگالی دو مدی متقارن است.*

**تعریف ۷.۳.** *فرض کنید $Z \sim SFN(\alpha,\delta)$ با $\alpha, \delta \in \mathbb{R}$. خانواده توزیع‌های چوله‌نرمال-انعطاف‌پذیر با پارامترهای مکان و مقیاس*

به‌صورت توزیع تبدیل $X = \mu + \sigma Z$ تعریف می‌شود، که در آن $\mu \in \mathbb{R}$ و $\sigma > 0$. تابع چگالی $X$ به‌صورت

$$f(x|\mu,\sigma,\alpha,\delta) = \frac{c_\delta}{\sigma}\phi(\frac{|x-\mu|}{\sigma}+\delta)\Phi(\alpha\frac{x-\mu}{\sigma}) \qquad (۱۵)$$

است. این توزیع به‌صورت $X \sim SFN(\theta)$ نمایش داده می‌شود که در آن $\theta = (\mu,\sigma,\alpha,\delta)$.

گشتاور اول تا چهارم برای $Z \sim SFN(\alpha,\delta)$ عبارت‌اند از:

$$E(Z) = -c_\delta[\delta\Phi(-\delta) - ۲\lambda\phi(\lambda\delta)\Phi(-\delta\lambda/\alpha)] + ۲\delta\mu_\circ(-\delta,\alpha,\delta)$$

$$E(Z^۲) = c_\delta[(۱+\delta^۲)\Phi(-\delta) - \delta\phi(\delta)]$$

$$E(Z^۳) = -c_\delta[\delta(۳+\delta)\Phi(-\delta) - ۲\lambda\phi(\lambda\delta)\{[۲+(۳+\lambda^۴-۳\lambda^۳)\delta^۲+\lambda^۲ \alpha^۲]\Phi(-\delta\lambda/\alpha) + [(\lambda^۳\delta/\alpha)(۲+۱/\alpha^۲) - ۳\lambda\delta/\alpha]\phi(\delta\lambda/\alpha)\}] + ۲\delta(۳+\delta^۲)\mu_\circ(-\delta,\alpha,\delta)$$

$$E(Z^۴) = c_\delta[(۳+۶\delta^۲+\delta^۴)\Phi(-\delta) - \delta(۵+\delta^۲)\phi(\delta)]$$

که در آن $\lambda = \frac{\alpha}{\sqrt{۱+\alpha^۲}}$ است و $\mu_\circ(-\delta,\alpha,\delta)$ باید به‌صورت عددی محاسبه شود(گومز و همکاران [۹]).

## ۲.۳ توزیع چوله‌نرمال دو مدی ($BSN$)

در این زیر بخش به بررسی توسعه دیگری از توزیع چوله‌نرمال پرداخته می‌شود که بر اساس فرع ۲.۳ ساخته‌شده است. این توزیع توسط اِلال-اُلیویرو و همکاران [۸] معرفی‌شده است. ابتدا توزیع نرمال دو مدی معرفی می‌شود.

**تعریف ۸.۳.** *اگر $X$ یک متغیر تصادفی با تابع چگالی*

$$f(x) = x^۲\phi(x); \qquad x \in \mathbb{R} \qquad (۱۶)$$

*باشد، که در آن $\phi$ تابع چگالی نرمال استاندارد، در این صورت $X$ دارای توزیع نرمال دو مدی نامیده می‌شود، که با نماد $X \sim BN$ نشان داده می‌شود (اِلال-اُلیویرو [۷]).*

**تذکر ۹.۳.** *اگر $X \sim BN$ با تابع چگالی $f(x)$، آنگاه $f(x)$ دو مدی است. این موضوع را می‌توان با توجه به اینکه $f'(x) = x\phi(x)(۲-x^۲)$ نتیجه گرفت، که کمترین مقدار تابع $f(x)$ در $x = ۰$، و بیشترین مقدار در $x = -\sqrt{۲}$ و $x = \sqrt{۲}$ است. توجه داشته باشید که $f(\sqrt{۲}) = f(-\sqrt{۲})$ است، این واقعیت با اضافه کردن یک پارامتر اضافی که ارتفاع را در مدها کنترل می‌کند، منجر به شکل‌گیری یک مدل انعطاف‌پذیر می‌شود که در ادامه مطرح خواهد شد.*

**تعریف ۱۰.۳.** *اگر $X$ یک متغیر تصادفی با تابع چگالی*

$$f(x|\delta) = (\frac{۱+\delta x^۲}{۱+\delta})\phi(x); \qquad x \in \mathbb{R}, \delta \geq ۰ \qquad (۱۷)$$

*باشد، در این صورت $X$ دارای توزیع نرمال دو مدی با پارامتر شکل $\delta$ است، که با نماد $X \sim BN(\delta)$ نشان داده می‌شود.*

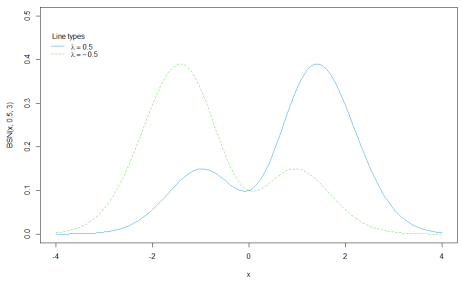
**گزارهٔ ۱۱.۳.** *فرض کنید $X$ و $Y$ متغیرهای تصادفی مستقل به‌گونه‌ای که $X \sim BN$ و $Y \sim N(۰,۱)$، اگر $W = \sqrt{\frac{\delta}{۱+\delta}}X + \sqrt{\frac{۱}{۱+\delta}}Y$ آنگاه $W \sim BN(\delta)$ است.*

**تعریف ۱۲.۳.** *اگر متغیر تصادفی $X$ دارای تابع چگالی*

$$f(x|\alpha,\delta) = ۲\Big(\frac{۱+\delta x^۲}{۱+\delta}\Big)\phi(x)\Phi(\alpha x); \qquad x \in \mathbb{R}, \alpha \in \mathbb{R}, \delta \geq ۰ \qquad (۱۸)$$

*آنگاه $X$ دارای توزیع چوله‌نرمال دو مدی با پارامترهای $\alpha$ و $\delta$، که با نماد $X \sim BSN(\alpha,\delta)$ نشان داده می‌شود. شکل ۲.۳ تابع چگالی معادله (۱۸) را به ازای مقادیر مختلف پارامترهای $\alpha$ و $\delta$ نشان می‌دهد. با توجه به این شکل $\alpha$ پارامتر چولگی و $\delta$ پارامتر کنترل‌کننده مد توزیع است، که به ازای مقادیر $\delta = ۰$ توزیع تک مدی است.*

لازم به ذکر است تابع چگالی (۱۸) به ازای مقادیر مختلف پارامترهای $\alpha$ و $\delta$ تغییر می‌کند، که نشان‌دهنده شناساپذیر بودن این توزیع است (گومز و همکاران [۹]).




**شکل ۲.۳.** تابع چگالی توزیع $BSN(\alpha,\delta)$ برای $\delta = ۳$ و مقادیر مختلف $\alpha$.

**گزارهٔ ۱۳.۳.** *اگر $X \sim BN(\alpha,\delta)$، آنگاه خواص زیر برقرار می‌باشند.*

- *اگر $\alpha = ۰$ آنگاه $X \sim BN(\delta)$*
- *اگر $\alpha = ۰$ و $\delta = ۰$ آنگاه $X \sim N(۰,۱)$*
- *اگر $\delta = ۰$ آنگاه $X \sim SN(\alpha)$*
- *اگر $\delta \to \infty$ آنگاه $X \sim BN(\alpha)$*
- *خاصیت ناوردایی، اگر $W \sim BSN(\alpha,\delta)$ و $Z \sim BN(\delta)$ آنگاه $|X| \stackrel{d}{=} |Z|$ هم توزیع می‌باشند.*

**تعریف ۱۴.۳.** *فرض کنید $Z \sim BSN(\alpha,\delta)$ با $\alpha,\delta \in \mathbb{R}$. خانواده توزیع‌های چوله‌نرمال دو مدی با پارامترهای مکان و مقیاس به‌صورت $X = \mu+\sigma Z$ تعریف می‌شود، که در آن $\mu \in \mathbb{R}$ و $\sigma > 0$. تابع چگالی $X$ به‌صورت*

$$f_X(x;\mu,\sigma,\alpha,\delta) = 2\frac{\sigma^2+\delta(x-\mu)^2}{\sigma^3(1+\delta)}\phi\Big(\frac{x-\mu}{\sigma}\Big)\times \Phi\Big(\alpha\frac{x-\mu}{\sigma}\Big), \qquad x\in\mathbb{R} \qquad (19)$$

*است. این توزیع به‌صورت $X \sim BSN(\theta)$ نمایش داده می‌شود که در آن $\theta = (\mu,\sigma,\alpha,\delta)$.*

**تعریف ۱۵.۳.** *فرض کنید $Z \sim BSN(\alpha,\delta)$ آنگاه گشتاورهای اول تا چهارم برای این توزیع عبارت‌اند از:*

$$E(X) = \frac{b\lambda}{(1+\delta)(1+\alpha^2)}[(1+2\alpha)\delta^2+3\alpha+1]$$
$$E(X^2) = \frac{1+3\alpha}{1+\alpha}$$
$$E(X^3) = \frac{b\lambda}{(1+\delta)(1+\alpha^2)^2}[(2+8\delta)\alpha^4+(5+20\delta)\alpha^2 + 15\delta+3]$$
$$E(X^4) = \frac{3+15\delta}{1+\delta}$$

*که در آن $b = 1/\sqrt{(1+\alpha^2)}$ و $\lambda = \sqrt{\frac{2}{\pi}}\alpha$ است.*

در این بخش توسعه‌های توزیع چوله‌نرمال به حالت دو مدی موردمطالعه قرار گرفت. در ادامه نحوه به‌کارگیری این توزیع‌ها در ساخت مدل‌های رگرسیونی منعطف ارائه می‌شود.

# ۴ مدل‌های رگرسیونی منعطف برای داده‌های نامتقارن دومدی

در این بخش مدل‌های رگرسیون مبتنی بر توزیع‌های $SFN$ و $BSN$ به‌عنوان یک فعالیت پژوهشی جدید معرفی می‌شوند.

## ۱.۴ مدل رگرسیون مبتنی بر توزیع $SFN$

**تعریف ۱.۴.** *یک مدل رگرسیون خطی با باقیمانده‌های $SFN$ که با نماد $SFN-LR$ نمایش داده می‌شود، به‌صورت*

$$Y_i = x_i^\top\beta+\epsilon_i, \qquad i=1,\cdots,n \qquad (20)$$

*معرفی می‌شود، که در آن $Y_i$ متغیر پاسخ، $x_i = (1,x_{i1},\cdots,x_{ip})^\top$ بردار مقادیر متغیر مستقل با بعد $1\times(p+1)$، $\beta = (\beta_0,\beta_1,\cdots,\beta_p)^\top$ بردار ضرایب رگرسیونی و خطای تصادفی $\epsilon_i \overset{iid}{\sim} SFN(0,\sigma,\alpha,\delta)$. در این صورت با توجه به خواص توزیع $SFN$، متغیر پاسخ $Y_i|x_i \overset{iid}{\sim} SFN(x_i^\top\beta,\sigma,\alpha,\delta)$ که تابع چگالی آن به‌صورت*

$$f_{Y_i}(y_i|x_i,\theta) = \frac{c_\delta}{\sigma}\phi\left(\frac{|y_i-x_i^\top\beta|}{\sigma}+\delta\right)\Phi\left(\alpha\frac{y_i-x_i^\top\beta}{\sigma}\right) \qquad (21)$$

*که در آن $\phi$ و $\Phi$ به ترتیب تابع چگالی و تابع توزیع نرمال استاندارد و $\theta = (\beta^\top,\sigma,\alpha,\delta)$ بردار پارامترهای مدل است.*

لگاریتم تابع درستنمایی با توجه به نمونه مشاهده‌شده $y_1,\cdots,y_n$ به‌صورت

$$l(\theta) = \sum_{i=1}^n l_i(\theta) = C - \frac{n}{2}\Big(\log(\frac{\sigma}{c_\delta})+\delta^2\Big) - \frac{\delta}{\sigma}\sum_{i=1}^n|y_i-x_i^\top\beta| - \frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-x_i^\top\beta)^2 + \sum_{i=1}^n\log\Phi\Big(\alpha\frac{y_i-x_i^\top\beta}{\sigma}\Big) \qquad (22)$$

است. با مشتق گرفتن از $l(\theta)$ نسبت به بردار پارامترهای مدل خواهیم داشت:

$$\frac{\partial l(\theta)}{\partial\beta_j} = \sum_{i=1}^n x_{ij}\Big(\frac{y_i-x_i^\top\beta}{\sigma^2}\Big)+\frac{\delta}{\sigma}\sum_{i=1}^n x_{ij}sign\Big(y_i-x_i^\top\beta\Big) - \frac{\alpha}{\sigma}\sum_{i=1}^n x_{ij}\eta\Big(\alpha\frac{y_i-x_i^\top\beta}{\sigma}\Big) = 0$$

$$\frac{\partial l(\theta)}{\partial\sigma} = -\frac{n}{2\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^n\Big(y_i-x_i^\top\beta\Big)^2-\frac{\alpha}{\sigma^2} \sum_{i=1}^n\eta\Big(\alpha\frac{y_i-x_i^\top\beta}{\sigma}\Big)\Big(y_i-x_i^\top\beta\Big)+\frac{\delta}{\sigma^2}\sum_{i=1}^n|y_i-x_i^\top\beta| = 0$$

$$\frac{\partial l(\theta)}{\partial\alpha} = \frac{1}{\sigma}\sum_{i=1}^n\eta\Big(\alpha\frac{y_i-x_i^\top\beta}{\sigma}\Big)\Big(y_i-x_i^\top\beta\Big) = 0$$

$$\frac{\partial l(\theta)}{\partial\delta} = \Big(c_\delta\phi(\delta)-\delta\Big)-\frac{1}{\sigma}\sum_{i=1}^n|y_i-x_i^\top\beta| = 0$$

که در آن $\eta(a) = \phi(a)/\Phi(a)$. بنابراین برآوردگر $ML$ از حل معادلات بالا به دست می‌آید، که می‌بایست روش‌های عددی مورداستفاده قرار گیرد.

## ۲.۴ مدل رگرسیون مبتنی بر توزیع $BSN$

**تعریف ۲.۴.** *یک مدل رگرسیون خطی با باقیمانده‌های $BSN$ که با نماد $BSN-LR$ نمایش داده می‌شود، به‌صورت*

$$Y_i = x_i^\top \beta + \epsilon_i, \qquad i = ۱, \cdots, n \qquad (۲۳)$$

*معرفی می‌شود، که در آن $Y_i$ متغیر پاسخ، $x_i = (۱, x_{i۱}, \cdots, x_{ip})^\top$ بردار مقادیر متغیر مستقل با بعد $(p+۱)\times ۱$، $\beta = (\beta_۰, \beta_۱, \cdots, \beta_p)^\top$ بردار ضرایب رگرسیونی و خطای تصادفی $\epsilon_i \overset{iid}{\sim} BSN(۰, \sigma, \alpha, \delta)$.*

در این صورت با توجه به خواص توزیع $BSN$، متغیر پاسخ $Y_i|x_i \overset{iid}{\sim} BSN(x_i^\top\beta, \sigma, \alpha, \delta)$ که تابع چگالی آن به‌صورت

$$f_{Y_i}(y_i|x_i,\theta) = ۲\frac{\sigma^۲ + \delta(y_i - x_i^\top\beta)^۲}{\sigma^۳(۱+\delta)}\phi\left(\frac{y_i - x_i^\top\beta}{\sigma}\right)\times \Phi\left(\alpha\frac{y_i - x_i^\top\beta}{\sigma}\right); \qquad y \in \mathbb{R} \qquad (۲۴)$$

که در آن $\phi$ و $\Phi$ به ترتیب تابع چگالی و تابع توزیع نرمال استاندارد و $\theta = (\beta^\top, \sigma, \delta, \alpha)$ بردار پارامترهای مدل است.

لگاریتم تابع درستنمایی با توجه به نمونه مشاهده‌شده $y_۱, \cdots, y_n$ به‌صورت

$$\begin{aligned} l(\theta) = \sum_{i=۱}^n l_i(\theta) &= -۳n\log(\sigma) - n\log(۱+\delta) \\ &\quad - \frac{۱}{۲\sigma^۲}\sum_{i=۱}^n (y_i - x_i^\top\beta)^۲ \\ &\quad + \sum_{i=۱}^n \log\left(\sigma^۲ + \delta(y_i - x_i^\top\beta\right) \\ &\quad + \sum_{i=۱}^n \log\Phi\left(\alpha\frac{y_i - x_i^\top\beta}{\sigma}\right) \end{aligned}$$

است. با مشتق گرفتن از $l(\theta)$ نسبت به بردار پارامترهای مدل خواهیم داشت:

$$\begin{aligned} \frac{\partial l(\theta)}{\partial\beta_j} &= \sum_{i=۱}^n x_{ij}\left(\frac{y_i - x_i^\top\beta}{\sigma^۲}\right) - ۲\delta x_{ij}\sum_{i=۱}^n \frac{(y_i - x_i^\top\beta)}{\sigma^۲ + \delta(y_i - x_i^\top\beta)^۲} \\ &\quad - \frac{۱}{\sigma^۲}\sum_{i=۱}^n x_{ij}\eta\left(\alpha\frac{y_i - x_i^\top\beta}{\sigma}\right) = ۰ \\ \frac{\partial l(\theta)}{\partial\sigma} &= -\frac{۳n}{\sigma} + \frac{۱}{\sigma^۳}\sum_{i=۱}^n (y_i - x_i^\top\beta)^۲ + ۲\sigma\sum_{i=۱}^n \frac{۱}{\sigma^۲ + \delta(y_i - x_i^\top\beta)^۲} \\ &\quad - \frac{۱}{\sigma^۲}\sum_{i=۱}^n (y_i - x_i^\top\beta)\eta\left(\alpha\frac{y_i - x_i^\top\beta}{\sigma}\right) = ۰ \\ \frac{\partial l(\theta)}{\partial\delta} &= -\frac{n}{۱+\delta} + \sum_{i=۱}^n \frac{(y_i - x_i^\top\beta)^۲}{\sigma^۲ + \delta(y_i - x_i^\top\beta)^۲} = ۰ \\ \frac{\partial l(\theta)}{\partial\alpha} &= \frac{۱}{\sigma}\sum_{i=۱}^n (y_i - x_i^\top\beta)\eta\left(\alpha\frac{y_i - x_i^\top\beta}{\sigma}\right) = ۰ \end{aligned}$$

که در آن $\eta(a) = \phi(a)/\Phi(a)$. بنابراین برآوردگر $ML$ از حل معادلات بالا به روش عددی به دست می‌آید.

در این راستا از الگوریتم نیلدر- مید در تابع *optim* موجود در نرم‌افزار $R$ استفاده می‌شود. این الگوریتم یکی از شناخته‌شده‌ترین روش‌های بدون مشتق است، که فقط از مقادیر تابع برای جستجوی مینیمم استفاده می‌کند. این رویکرد شامل ساخت یک سیمپلکس از $n+۱$ نقطه و حرکت این سیمپلکس در جهت صحیح است. برای آشنایی بیشتر با این الگوریتم به این سایت مراجعه کنید.

# ۵ کاربرد و ارزیابی مدل‌های رگرسیونی منعطف

در این بخش با استفاده از شبیه‌سازی و مثال کاربردی مربوط به مجموعه داده‌های اسب‌دوانی $(horses)$، نحوه کاربست مدل‌های رگرسیونی معرفی‌شده در این مقاله مورد ارزیابی قرار می‌گیرد.

## ۱.۵ مطالعه شبیه‌سازی

در این زیر بخش، شبیه‌سازی مختصری از مدل‌های رگرسیونی $BSN-LR$ و $SFN-LR$ برای بررسی رفتار برآوردگرها در نمونه‌های متناهی ارائه می‌شود. برای محاسبه برآوردگرهای $MLE$ از الگوریتم $Nelder-Mead$ در تابع *optim* موجود در نرم‌افزار $R$ استفاده می‌شود. شبیه‌سازی از مدل رگرسیونی

$$Y_i = ۲ + ۵X_{۱i} + ۳X_{۲i} + \epsilon_i \qquad i = ۱, \cdots, n$$

صورت می‌گیرد، که در آن ابتدا خطای تصادفی دارای توزیع $\epsilon_i \sim BSN(۰, ۴, -۰٫۵, ۳)$ و بار دیگر $\epsilon_i \sim SFN(۰, ۴, ۰٫۵, -۳)$ است. حجم نمونه‌ها ۱۰۰، ۵۰۰ و ۱۰۰۰ می‌باشد. همان‌طور که از جداول ۱ و ۲ مشاهده می‌شود، بر اساس ویژگی‌های مجانبی برآوردگرهای $MLE$ با افزایش حجم نمونه مقدار اریبی برآوردها کمتر می‌شود، مقدار خطای برآوردگرها و همچنین $AIC$ مدل نیز کاهش می‌یابد. در ادامه این زیر بخش، با استفاده از شبیه‌سازی، مدل‌های $SFN-LR$ و $BSN-LR$ با $Mix.SN-LR$ مورد مقایسه قرار می‌گیرند. برای محاسبه برآوردگرهای $MLE$ مدل‌های $SFN-LR$ و $BSN-LR$ از تابع *optim* و همچنین برای برازش مدل رگرسیون آمیخته‌ای از دو توزیع چوله‌نرمال از کتابخانه $FMsmsnReg$ موجود در نرم‌افزار $R$ استفاده می‌شود. ابتدا، خطای تصادفی مدل با استفاده از آمیخته‌ای از

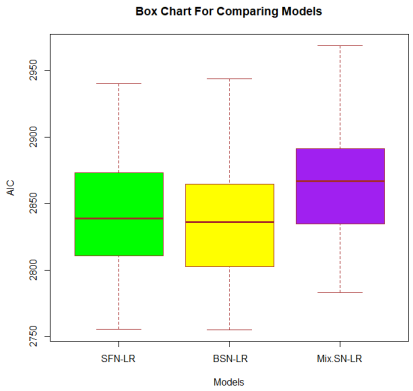
دو توزیع چوله‌نرمال $\epsilon \sim Mix.SN(۰, ۰, ۱, ۱, ۰/۲, ۰/۲, ۰/۳)$ نمونه‌ای تصادفی به حجم ۱۰۰۰ با تکرار ۵۰ تولید می‌شود، سپس مدل رگرسیونی

$$Y_i = ۴ + ۲X_{۱i} - ۲X_{۲i} + \epsilon_i \qquad i = ۱, \cdots, n$$

شبیه‌سازی می‌شود، در ادامه مدل‌های رگرسیونی معرفی‌شده یعنی $SFN - LR$ و $BSN - LR$ با $Mix.SN - LR$ به داده‌ها برازش می‌شود و مقادیر $AIC$، $Bias(\hat{\beta})$ و $MSE(\hat{\beta})$ به‌صورت

$$Bias(\hat{\beta}) = (\bar{\hat{\beta}}) - \beta_{True}, \qquad MSE(\hat{\beta}) = \frac{۱}{۵۰}\sum_{i=۱}^{۵۰}(\hat{\beta} - \beta_{True})^۲$$

محاسبه می‌شود. نتایج در جدول ۳ ارائه‌شده است. در انتها نمودار جعبه‌ای برای مقادیر $AIC$ تحت مدل‌های رگرسیونی مختلف رسم می‌شود (شکل ۱.۵ را ملاحظه کنید). با توجه به این شکل و جدول ۳ مشاهده می‌شود بااینکه مجموعه داده‌ها از مدل $Mix.SN - LR$ تولیدشده است، عملکرد مدل $BSN - LR$ نسبت به سایر مدل‌ها مناسب‌تر است. این امر می‌تواند به دلیل انعطاف‌پذیری بیشتر و تعداد پارامترهای کمتر مدل‌های چوله‌نرمال دو مدی نسبت به مدل‌های آمیخته‌ای از دو توزیع چوله‌نرمال باشد.




**شکل ۱.۵.** نمودار جعبه‌ای مقادیر $AIC$ تحت مدل‌های مختلف برای ۵۰ تکرار شبیه‌سازی.

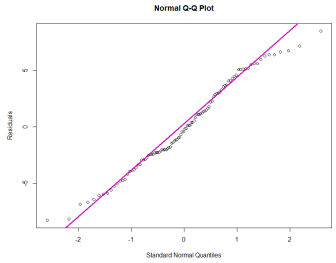
## ۲.۵ مثال کاربردی

در این زیر بخش به‌عنوان یک مثال انگیزشی، مجموعه داده‌ای را که قبلاً توسط فوربز (۱۹۹۸) به‌وسیله یک مدل رگرسیون نرمال تجزیه و تحلیل‌شده است، در نظر گرفته می‌شود. مجموعه داده‌های اسب‌دوانی $(horses)$، شامل نتایج مربوط به هر اسب در ۸ مسابقه متوالی می‌باشد. داده‌ها شامل ۱۴ متغیر که هرکدام ۱۰۲ بار مشاهده‌شده است. در اینجا، متغیر پاسخ موقعیت پایان $(Position)$ و متغیرهای کمکی شامل تعداد اسب‌های مسابقه $(Starters)$ و نسبت بردها در شروع قبلی $(Ratio)$ هستند. بنابراین، مدل رگرسیونی

$$Position_i = \beta_۰ + \beta_۱ Starters_i + \beta_۲ Ratio_i + \epsilon_i$$

در نظر گرفته می‌شود. ابتدا، مدل رگرسیون خطی با دو متغیر مستقل به روش $OLS$ با دستور $lm$ به داده‌ها برازش داده می‌شود، در ادامه برای بررسی فرض نرمال بودن، باقیمانده‌های مدل محاسبه می‌شود. حال برای بررسی این فرض به‌صورت شهودی از نمودار چندک-چندک $(Q - Qplot)$، و در ادامه برای تحلیل دقیق‌تر از آزمون کولموگروف اسمیرنوف بر روی باقیمانده‌های مدل استفاده می‌شود.

نمودار چندک-چندک یا $Q - Qplot$ به‌منظور مقایسه دو توزیع به کار گرفته می‌شود. از چنین نمودارهایی حتی می‌توان مطابقت توزیع داده‌ها را با یک توزیع مشخص، موردبررسی قرار داد. نمودار چندک-چندک بر اساس چندک‌های دو توزیع ترسیم می‌شود. در محور افقی چندک‌های توزیع اول و در محور عمودی نیز چندک‌های متناظر برای توزیع دوم مشخص می‌شود. اگر این مقادیر را در یک صفحه مختصات دکارتی ترسیم کنیم، یک نمودار چندک چندک یا $Q - Qplot$ ایجاد کرده‌ایم.




**شکل ۲.۵.** نمودار چندک-چندک مربوط به باقیمانده‌های مدل $OLS$.

**جدول ۴.** نتایج آزمون کولموگروف-اسمیرنف بر روی باقیمانده‌های مدل $OLS$.

| آزمون | $p - value$ |
|---|---|
| کولموگروف-اسمیرنف | $۳/۹۰۴e^{-۱۲}$ |

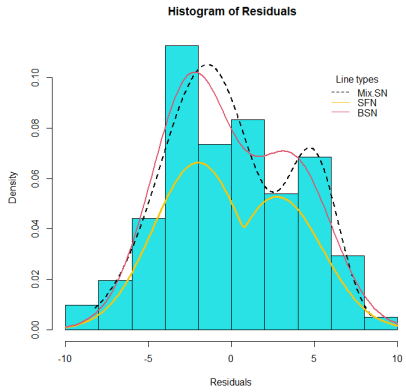
با توجه به شکل ۲.۵ مشاهده می‌شود، که باقیمانده‌ها به‌طور کامل روی خط قرار نمی‌گیرند، و از طرفی پی-مقدار مربوط به آزمون کولموگروف-اسمیرنف از سطح معنی‌داری ۰/۰۵ کمتر است (با توجه به جدول ۲.۵)، پس می‌توان نتیجه گرفت که فرض نرمال بودن باقیمانده‌های مدل برقرار نمی‌باشد. در ادامه هیستوگرام باقیمانده‌ها رسم می‌شود و توزیع‌های معرفی‌شده در این مقاله، بر روی آن برازش شده است. با توجه به شکل ۲.۵، هیستوگرام باقیمانده‌ها تک مدی و متقارن نمی‌باشد، بنابراین می‌توان از توزیع‌های $Mix.SN$، $SFN$

و $BSN$ برای مدل‌سازی این داده‌ها استفاده کرد. در جدول ۴ خلاصه برازش مدل‌های رگرسیونی به داده‌ها و برآورد پارامترهای مدل آورده شده است. به‌عنوان یک شاخص سنجش و انتخاب مدل مناسب، از معیار ارزیابی $AIC$ استفاده‌شده است. شاخص $AIC$ توسط دانشمند ژاپنی آمار، «هیروتاگا آکایکه» ($HirotuguAkaike$) در سال‌های ۱۹۷۰ برای تشخیص مدل مناسب از بین مدل‌های موجود، معرفی شد و امروز به‌عنوان یک ابزار مهم در انتخاب مدل‌ها بر اساس تابع درستنمایی به کار گرفته می‌شود.

فرض کنید در یک مدل آماری $k$ تعداد پارامترهای مدل باشد. اگر $\hat{L}$ را ماکسیمم تابع درستنمایی تحت مدل مفروض در نظر بگیریم، معیار ارزیابی $AIC$ توسط رابطه

$$AIC = ۲k - ۲\log \hat{L}$$

قابل‌محاسبه است. بنابراین مناسب‌ترین مدل برحسب معیار اطلاع آکایکه، دارای کمترین مقدار $AIC$ است. با توجه به اینکه مقدار $AIC$ برای مدل رگرسیونی مبتنی بر توزیع $BSN$ نسبت به مدل‌های دیگر کمتر است، در نتیجه این مدل برای برازش به داده‌ها عملکرد مناسب‌تری دارد.




**شکل ۳.۵.** هیستوگرام باقیمانده‌ها همراه با برازش توابع چگالی توزیع‌های معرفی‌شده.

## ۶ بحث و نتیجه‌گیری

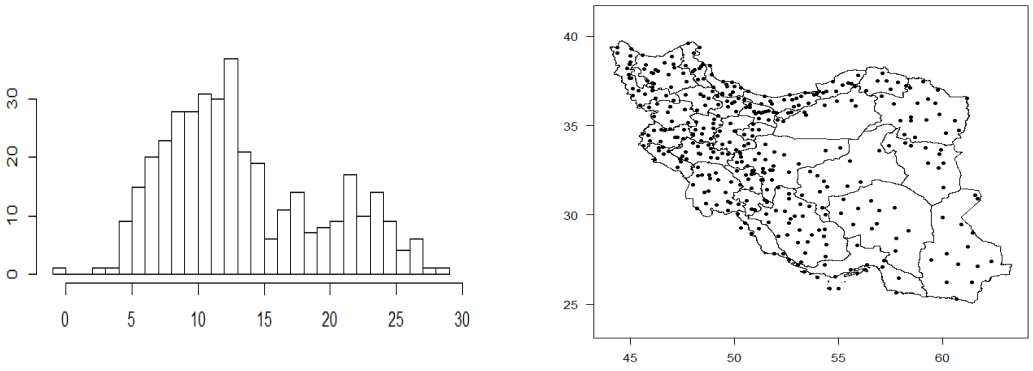
در این مقاله، به بررسی رویکردهای مختلف در تحلیل رگرسیونی برای حالتی که داده‌ها به‌طور هم‌زمان چولگی و دو مدی بودن را نشان می‌دهند، پرداخته شد. به‌ویژه مدل‌های رگرسیونی مبتنی بر توزیع‌های چوله‌نرمال دو مدی معرفی گردید، که این‌گونه مدل‌ها می‌توانند به‌عنوان جایگزینی برای مدل‌های آمیخته به کار روند. با استفاده از مطالعات شبیه‌سازی و مثال کاربردی، عملکرد مناسب‌تر این‌گونه مدل‌ها نشان داده شد. بااین‌حال در بسیاری از مسائل کاربردی ممکن است با داده‌هایی که ساختار همبستگی فضایی دارند مواجه شویم، در نتیجه فرض استقلال نیز برای این داده‌ها برقرار نمی‌باشد. در این حالت توسعه مدل‌های پیشنهادی به حالت فضایی به‌عنوان کار آینده مورد توجه نویسنده مقاله قرار دارد.

جدول ۱: نتایج شبیه‌سازی از مدل رگرسیونی $BSN - LR$.

| | $n = ۱۰۰$ | | | | $n = ۵۰۰$ | | | | $n = ۱۰۰۰$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | فاصله اطمینان ۹۵٪ | | | | فاصله اطمینان ۹۵٪ | | | | فاصله اطمینان ۹۵٪ | | |
| پارامترها | $MLE$ | خطای برآورد | پایین | بالا | برآورد $MLE$ | خطای برآورد | پایین | بالا | برآورد $MLE$ | خطای برآورد | پایین | بالا |
| $\beta_۰$ | ۱٫۵۴۷۱ | ۰٫۷۶۹۸ | ۰٫۰۳۸۳ | ۳٫۰۵۵۹ | ۲٫۲۴۶۲ | ۰٫۳۱۱۸ | ۱٫۶۳۵۱ | ۲٫۸۵۷۴ | ۲٫۲۳۴۹ | ۰٫۱۶۵۸ | ۱٫۹۰۹۹ | ۲٫۵۵۹۹ |
| $\beta_۱$ | ۴٫۶۶۲۴ | ۰٫۴۱۰۲ | ۳٫۸۵۸۴ | ۵٫۴۶۶۴ | ۵٫۰۲۲۵ | ۰٫۱۷۹۰ | ۴٫۶۷۱۶ | ۵٫۳۷۳۴ | ۵٫۰۰۱۲ | ۰٫۱۱۴۵ | ۴٫۷۷۶۸ | ۵٫۲۲۵۶ |
| $\beta_۲$ | ۲٫۹۵۵۲ | ۰٫۴۰۲۵ | ۲٫۱۶۶۲ | ۳٫۷۴۴۱ | ۲٫۹۹۱۵ | ۰٫۱۸۰۸ | ۲٫۶۳۷۰ | ۳٫۳۴۶۱ | ۳٫۰۶۱۱ | ۰٫۱۱۴۱ | ۲٫۸۳۷۵ | ۳٫۲۸۴۸ |
| $\sigma$ | ۳٫۰۱۹۰ | ۰٫۳۳۶۸ | ۲٫۳۵۸۷ | ۳٫۶۷۹۲ | ۳٫۱۸۱۳ | ۰٫۱۴۳۱ | ۲٫۹۰۰۷ | ۳٫۴۶۱۹ | ۳٫۹۰۷۸ | ۰٫۰۸۳۰ | ۳٫۷۴۶۹ | ۴٫۰۶۸۷ |
| $\delta$ | ۱٫۶۳۹۷ | ۰٫۷۵۱۳ | ۰٫۱۶۷۱ | ۳٫۱۱۲۳ | ۲٫۷۱۵۰ | ۰٫۵۹۴۳ | ۱٫۵۵۰۲ | ۳٫۸۷۹۸ | ۳٫۰۱۲۳ | ۰٫۱۷۱۴ | ۲٫۲۵۳۵ | ۳٫۵۴۲۷ |
| $\alpha$ | −۰٫۴۱۳۲ | ۰٫۱۵۹۸ | −۰٫۷۲۶۵ | −۰٫۱۰۰۱ | −۰٫۵۷۷۴ | ۰٫۰۶۹۰ | −۰٫۷۱۲۷ | −۰٫۴۴۲۱ | −۰٫۵۰۰۵ | ۰٫۰۴۰۷ | −۰٫۵۸۰۴ | −۰٫۴۲۰۷ |
| $AIC$ | | | −۵۹۹٫۳۸۷۲ | | | | −۲۹۹٫۰۱۷۴ | | | | −۵۹۲٫۵۸۳۹ | |

## مراجع

[۱] بهرامی، م. (۱۳۹۴)، مدل‌های آمیخته متناهی مبتنی بر توزیع‌های چوله متقارن، پایان نامه دوره ارشد، دانشگاه تربیت مدرس.




(ب) (الف)

شکل ۱: (الف) موقعیت مکانی ۳۸۱ ایستگاه هواشناسی در ایران و (ب) هیستوگرام متغیر حداقل دمای هوا.

جدول ۲: نتایج شبیه‌سازی از مدل رگرسیونی $SFN - LR$.

| | | $n = ۱۰۰$ | | | | $n = ۵۰۰$ | | | | $n = ۱۰۰۰$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | فاصله اطمینان ۹۵٪ | | | | فاصله اطمینان ۹۵٪ | | | | فاصله اطمینان ۹۵٪ | | | |
| پارامترها | $MLE$ | خطای برآورد | پایین | بالا | برآورد $MLE$ | خطای برآورد | پایین | بالا | برآورد $MLE$ | خطای برآورد | پایین | بالا |
| $\beta_0$ | ۱٫۲۷۵۰ | ۰٫۷۴۳۱ | −۰٫۱۸۱۶ | ۲٫۳۱۴ | ۱٫۸۶۶۳ | ۰٫۲۹۰۹ | ۱٫۳۹۶۰ | ۲٫۵۳۶۶ | ۱٫۹۴۷۸ | ۰٫۲۱۹۵ | ۱٫۵۱۷۴ | ۲٫۳۷۸۱ |
| $\beta_1$ | ۵٫۳۴۷۰ | ۰٫۳۹۱۶ | ۴٫۵۷۹۵ | ۶٫۱۱۴۶ | ۴٫۹۷۶۱ | ۰٫۱۶۶۶ | ۴٫۶۴۹۴ | ۵٫۳۰۲۸ | ۵٫۱۰۳۴ | ۰٫۱۲۶۱ | ۴٫۸۵۶۱ | ۵٫۳۵۰۷ |
| $\beta_2$ | ۲٫۳۵۲۵ | ۰٫۴۳۴۲ | ۱٫۵۰۲۴ | ۳٫۲۰۴۶ | ۳٫۰۶۹۷ | ۰٫۱۷۸۷ | ۲٫۷۱۹۳ | ۳٫۴۲۰۱ | ۳٫۰۸۰۳ | ۰٫۱۱۸۱ | ۲٫۸۴۸۷ | ۳٫۳۱۲۰ |
| $\sigma$ | ۴٫۲۲۹۹ | ۰٫۳۲۷۱ | ۳٫۵۸۸۷ | ۴٫۸۷۱۱ | ۳٫۸۹۳۵ | ۰٫۱۳۲۵ | ۳٫۶۳۳۸ | ۴٫۱۵۳۲ | ۴٫۰۵۵۸ | ۰٫۰۹۸۳ | ۳٫۸۶۳۰ | ۴٫۲۴۸۶ |
| $\alpha$ | ۰٫۵۵۹۶ | ۰٫۰۹۲۷ | ۰٫۳۷۷۹ | ۰٫۷۴۱۳ | ۰٫۴۶۶۲ | ۰٫۰۳۳۹ | ۰٫۳۹۹۶ | ۰٫۵۳۲۸ | ۰٫۵۰۳۰ | ۰٫۰۲۶۰ | ۰٫۴۵۲۰ | ۰٫۵۵۴۰ |
| $\delta$ | −۲٫۸۹۹۹ | ۰٫۲۵۴۱ | −۳٫۳۹۸۱ | −۲٫۴۰۱۷ | −۳٫۱۰۶۸ | ۰٫۱۱۵۲ | −۳٫۳۳۳۰ | −۲٫۸۸۰۶ | −۳٫۰۰۰۴ | ۰٫۰۸۰۷ | −۳٫۱۵۸۷ | −۲٫۸۴۲۰ |
| $AIC$ | | | −۶۰۶٫۰۱۹۷ | | | | −۳۰۴۱٫۸۸۹ | | | | −۶۱۳۸٫۰۸۳ | |

جدول ۳: خلاصه برازش مدل‌های رگرسیونی به داده‌های شبیه‌سازی شده.

| | | $BSN - LR$ | | | $SFN - LR$ | | | $Mix.SN - LR$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| پارامترها | مقدار واقعی پارامتر | $MLE$ | $Bias(\hat{\beta})$ | $MSE(\hat{\beta})$ | $MLE$ | $Bias(\hat{\beta})$ | $MSE(\hat{\beta})$ | $MLE$ | $Bias(\hat{\beta})$ | $MSE(\hat{\beta})$ |
| $\beta_0$ | ۴ | ۲٫۷۰۵۱ | −۱٫۲۹۴۸ | ۱٫۷۱۳۸ | ۳٫۷۳۷۷ | −۰٫۲۶۲۲ | ۰٫۳۷۹۵ | ۴٫۱۵۶۲ | ۰٫۱۵۶۲ | ۰٫۰۲۵۰ |
| $\beta_1$ | ۲ | ۲٫۰۰۲۹ | ۰٫۰۰۲۹ | ۰٫۰۰۱۰ | ۲٫۰۰۱۷ | ۰٫۰۰۱۷ | ۰٫۰۰۱۳ | ۱٫۹۹۹۸ | −۰٫۰۰۰۲ | ۰٫۰۰۱۰ |
| $\beta_2$ | −۲ | −۱٫۹۹۷۶ | ۰٫۰۰۲۳ | ۰٫۰۰۰۷ | −۱٫۹۹۴۴ | ۰٫۰۰۵۵ | ۰٫۰۰۱۴ | −۱٫۹۹۷۷ | ۰٫۰۰۲۲ | ۰٫۰۰۰۷ |
| $AIC$ | | ۲۸۳۷٫۴۵۹ | | | ۲۸۴۳٫۸۸۱ | | | ۲۸۶۷٫۳۶۹ | | |


[2] Arellano-Valle, R. B., Gómez, H. W., and Quintana, F. A. (2004). A New Class of Skew-Normal Distributions. *Communications in Statistics-Theory and Methods*, **33**(7), 1465–1480.

[3] Arellano-Valle, R. B., Castro, L. M., Genton, M. G., and Gómez, H. W. (2008). Bayesian Inference for Shape Mixtures of Skewed Distributions, with Application to Regression Analysis. *Bayesian Analysis*, **3**(3), 513-539.

[4] Azzalini, A. (1985). A Class of Distributions Which Includes the Normal Ones. *Scandinavian journal of statistics*, **12**, 171-178.

[5] Azzalini, A., and Capitanio, A. (1999). Statistical Applications of the Multivariate Skew Normal Distribution. *Journal of the Royal Statistical Society*, Series B, **61**(3), 579-602.

جدول ۴: نتایج برازش مدل‌های رگرسیونی در نظر گرفته‌شده به داده‌ها.

| $BSN-LR$ | | $SFN-LR$ | | $Mix.SN-LR$ | | |
|---|---|---|---|---|---|---|
| خطای برآورد | برآورد $MLE$ | خطای برآورد | برآورد $MLE$ | خطای برآورد | $MLE$ | پارامترها |
| ۲٫۰۷ | ۵٫۶۰ | ۱٫۳۱ | ۰٫۹۶ | ۰٫۰۰۹ | ۲٫۰۴ | $\beta_0$ |
| ۰٫۱۱ | ۰٫۳۰ | ۰٫۰۹ | ۰٫۴۸ | ۰٫۱۵ | ۰٫۵۰۴ | $\beta_1$ |
| ۳٫۲۲ | −۶٫۱۰ | ۲٫۹۱ | −۰٫۵۲ | ۰٫۲۳ | −۸٫۴۳ | $\beta_2$ |
| – | – | – | – | ۳٫۹۳ | ۳٫۷۰ | $\mu_1$ |
| – | – | – | – | ۲٫۱۷ | −۲٫۶۳ | $\mu_2$ |
| ۰٫۲۳ | ۲٫۸۲ | ۰٫۳۸ | ۲٫۸۲ | ۱٫۸۰ | ۶٫۵۲ | $\sigma_1$ |
| – | – | – | – | ۳٫۳۴ | ۹٫۵۱ | $\sigma_2$ |
| ۰٫۱۹ | −۰٫۲۶ | ۰٫۱۲ | −۰٫۰۲ | ۶٫۰۷ | ۰٫۷۴ | $\alpha_1$ |
| – | – | – | – | ۱٫۱۲ | −۱٫۳۲ | $\alpha_2$ |
| ۰٫۷۲ | ۱٫۴۳ | ۰٫۳۰ | −۰٫۸۵ | – | – | $\delta$ |
| – | – | – | – | ۱٫۲۰ | ۰٫۴۱ | $\eta$ |
| | ۵۵۰٫۷۰۷۹ | | ۵۵۴٫۴۰۱۶ | | ۵۷۸٫۲۰۵ | $AIC$ |


[6] Cancho, V. G., 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**(1), 588-602.

[7] Elal-Olivero, D. (2010), Alpha-Skew-Normal Distribution. *Proyecciones Journal of Mathematics*. **29**(3), 224–240., 11, 12.

[8] Elal-Olivero, D., Olivares-Pacheco, J. F., Venegas, O., Bolfarine, H., and Gomez, H. W. (2020). On Properties of the Bimodal Skew-Normal Distribution and an Application. *Mathematics*, **8**(5), 703.

[9] Gómez, H. W., Elal-Olivero, D., Salinas, H. S., and Bolfarine, H. (2011). Bimodal Extension Based on the Skew-Normal Distribution with Application to Pollen Data. *Environmetrics*, **22**(1), 50-62.

[10] Henze, N. (1986). A Probabilistic Representation of the Skew-Normal Distribution, *Scandinavian Journal of Statistics*. **13**, 271-275.

[11] Lachos, V., Bolfarine, H., Arellano-Valle, R. and Montenegro, L. (2007). Likelihood Based Inference for Multivariate Skew-Normal Regression Models. *Communications in Statistics-Theory and Methods*, **36**, 1769–1786.

[12] Lin, T. I., Lee, J. C., and Yen, S. Y. (2007). Finite Mixture Modelling Using the Skew-Normal Distribution. *Statistica Sinica*, **17**: 81–92.

[13] Marin, J. M., Mengersen, K., and Robert, C. P. (2005). Bayesian Modelling and Inference on Mixtures of Distributions. *Handbook of Statistics*, **25**, 459-507.

[14] McLachan, G. and Peel, D. (2000). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, N.Y.

[15] Pewsey, A. (2003), The Characteristic Functions of the Skew-Normal and Wrapped Skew-Normal Distributions, In 27 Congreso Nacional de Estadisticae InvestigacionOperativa, pp. 4383-4386.

[16] Sartori, N. 2006. Bias Prevention of Maximum Likelihood Estimates for Scalar Skew-Normal and Skew-t Distributions. *J. Statist. Plann. Inference*, **136**, 4259–4275.

## پیوست:

تولید نمونه از توزیع $BSN$:

```
BSN=function(n,alpha,lambda){
x1=rchisq(n,3)
x2=rnorm(n)
u=c();w=c();t=c();s=c()
for (i in 1:n) {
u[i]=sample(c(-1,1),size = 1,prob = c(1/2,1/2))
w[i]=sqrt(x1[i])*u[i]
t[i]=sqrt(alpha/(1+lambda))*w[i]+sqrt(1/(1+lambda))*x2[i]
s[i]=sample(c(t[i],-t[i]),size = 1,prob = c(pnorm(alpha*t[i]),1-pnorm(alpha*t[i])))
}
return(s)
}
```

تولید نمونه از توزیع $SFN$:

```
library(TruncatedNormal)
SFN=function(n,alpha,delta){
y1=c();w1=c() ; s1=c() ; z=c()
for(i in 1:n){
y1[i]=rtnorm(1,-delta,lb=0,ub=Inf,sd=1)
w1[i]=pnorm(alpha*y1[i])
s1[i]=sample(c(1,-1),size =1,replace = TRUE,prob = c(w1[i],1-w1[i]))
z[i]=s1[i]*y1[i]
}
return(z)
}
```

برازش مدل رگرسیون به روش $OLS$ و بررسی نرمال بودن باقیمانده‌های مدل:

```
library(FMsmsnReg)
library(e1071)
data("horses")
head(horses)
str(horses)
attach(horses)
model.ols=lm(Position~Starters+Ratio)
skewness(model.ols$re)
```

```
kurtosis(model.ols$re)
sd(model.ols$re)
mean(model.ols$re)
##############################
hist(model.ols$res,breaks = 10,col=13,main= "Residuals of ols method",xlab = "Ordinary
    residuals",freq = FALSE)
ks.test(model.ols$res,"pnorm")
qqnorm(model.ols$res,ylab="Residuals",xlab = "Standard Normal Quantiles")
qqline(model.ols$res,col=6,lwd=3)
```

برازش مدل‌های رگرسیونی معرفی‌شده در بخش ۴:

```
#################### BSN-LR
regression.BSN=function(y,x1,x2,beta){
b0=beta[1]
b1=beta[2]
b2=beta[3]
alpha=beta[4]
lambda=beta[5]
sigma=beta[6]
lhod=2*(((sigma^2)+lambda*(y-b0-b1*x1-b2*x2)^2)/(sigma^3*(1+lambda)))*dnorm((y-b0-b1*x1-b2
    *x2)/sigma)*pnorm(alpha*(y-b0-b1*x1-b2*x2)/sigma)
return(-sum(log(lhod)))
}
MLE.BSN=optim(c(3,2,-1,1,-1,2),regression.BSN,y=Position,x1=Starters,x2=Ratio,hessian =
    TRUE)
MLE.BS=MLE.BSN$par
ObsInfo=MLE.BSN$hess #Observed Fisher information matrix
Vhat=solve(ObsInfo) #Inverse of observed Fisher information
Std.Errors=sqrt(diag(Vhat))
#Obtain the MLEs,estimated std errors, and approx Wald 95% Cis
Wald.table.BSN=cbind(MLE.BS,Std.Errors,LowerBound=MLE.BS-qnorm(0.975)*Std.Errors,
UpperBound=MLE.BS+qnorm(0.975)*Std.Errors)
Wald.table.BSN
row.names(Wald.table.BSN )=c("beta0","beta1","beta2","alpha","delta","sigma")
Wald.table.BSN
K=6
AIC=2*K-2* MLE.BSN$value
```

```
natijeh.BSN=list(Wald.table.BSN,AIC=AIC)
################################### SFN-LR
regression.SFN=function(beta,y,x1,x2){
b0=beta[1]
b1=beta[2]; b2=beta[3]; alpha=beta[4];delt=beta[5] ;sigma=beta[6];cdeta=1/(1-pnorm(delt))
lhd= (cdeta/sigma)*dnorm((abs(y-b0-b1*x1-b2*x2)/sigma)+delt)*pnorm(alpha*(y-b0-b1*x1-b2*x2
    )/sigma)
return(-sum(log(lhd)))
}
MLE.SFN=optim(c(3,2,-1,1,-1,2),regression.SFN,y=Position,x1=Starters,x2=Ratio,hessian =
    TRUE)
MLE.SF=MLE.SFN$par
ObsInfo=MLE.SFN$hess #Observed Fisher information matrix
Vhat=solve(ObsInfo) #Inverse of observed Fisher information
Std.Errors=sqrt(diag(Vhat))
#Obtain the MLEs,estimated std errors, and approx Wald 95% Cis
Wald.table.SFN=cbind(MLE.SF,Std.Errors,LowerBound=MLE.SF-qnorm(0.975)*Std.Errors,
UpperBound=MLE.SF+qnorm(0.975)*Std.Errors)
Wald.table.SFN
row.names(Wald.table.SFN)=c("beta0","beta1","beta2","alpha","delta","sigma")
Wald.table.SFN
K=6
AIC=2*K-2*MLE.SFN$value
natijeh.SFN=list(Wald.table.SFN,AIC=AIC)
######################################### Mix.sn.LR
data("horses")
head(horses)
attach(horses)
FMsmsnReg()
x=cbind(1,Starters,Ratio)
parCN <- FMsmsnReg( Position,x, g=2, get.init = TRUE, criteria = TRUE, group = FALSE,shape
    =c(0,0),
family = "Skew.normal", error = 10^-4,obs.prob= FALSE)
```

# Regression Models for Analyzing of Bimodal and Skewed Data

**Hassan Mirzavand[1], and Majid Jafari Khaledi[2]**

Abstract:

For statistical inference about the parameters of the regression model, it is necessary to assume a specific distribution for the random error term. A basic assumption in the linear regression model is that the random error term follows a normal distribution. However, in statistical research, sometimes the distribution of the data display both skewness and bimodality, and in such situations, it is inappropriate to use the normal distribution for statistical analysis. A conventional approach to overcome this problem is to use a mixture of normal models. But in such models, the number of parameters increases substantially, which makes it difficult to fit these models to the data. In addition, the mixed models suffer from the non-identifiability issues. In this case, a suitable solution is to use flexible distributions which can simultaneously handle the skewness and bimodality of the data in the modeling structure. So far, various methods have been proposed, which were created based on the development of the skew-normal distribution. In this article, these asymmetric bimodal distributions are used to build and introduce a flexible regression model compared to the regression models based on the normal distribution as well as a mixture of two normal distributions. Their performance is evaluated using a simulation example. Then, the usefulness of the method is demonstrated through a practical example related to the horse data set.

**Keywords:** Skewness, Bimodal distributions, Symmetric, Mixture of distributions, Regression .

[1]**Tarbiat Modares University, Tehran**

[2]**Tarbiat Modares University, Tehran.**