مجله مدل‌سازی پیشرفته ریاضی
سال ۱۴۰۴، دوره ۱۵، شماره ۳، ص ۱۲۷–۱۳۶

شاپا چاپی: ۸۰۸۸–۲۲۵۱
DOI: 10.22055/jamm.2025.49260.2362

رگرسیون چندکی برپایه توزیع‌های آمیخته مقیاس نرمال دوبخشی

محسن ملکی [1] [۱] و علیرضا نعمت اللهی [2]

[1] گروه آمار، دانشکده ریاضی و آمار، دانشگاه اصفهان، اصفهان ۷۳۴۴۱–۸۱۷۴۶، ایران
[2] گروه آمار، دانشکده علوم، دانشگاه شیراز، شیراز، ایران

دبیر مسئول: حمزه ترابی

تاریخ دریافت: ۱۴۰۴/۰۱/۱۳ تاریخ پذیرش: ۱۴۰۴/۰۸/۲۴

چکیده: مدل‌های رگرسیونی در برخی موارد بدلیل وجود مشاهدات دورافتاده و رفتار نامتقارن آنها نمی‌توانند برازش مناسبی روی داده‌ها داشته باشند، از این رو به کارگیری مدل‌های متقارن– نامتقارن دم سبک– سنگین در ساختار آنها می‌تواند مشکلات فوق را مرتفع سازد. در این مقاله یک مدل رگرسیون چندکی استوار با خطاهای توزیع شده برپایه‌ی توزیع‌های آمیخته– مقیاس نرمال دوبخشی که توسط ملکی و همکاران (۲۰۱۹) معرفی گردید را در راستای حل مسائل فوق معرفی خواهیم کرد و برآورد پارامترهای مدل را با رویکرد بیشینه درستنمایی بدست خواهیم آورد. خانواده‌ی آمیخته– مقیاس نرمال دوبخشی شامل توزیع‌های نرمال دوبخشی، تی دوبخشی، اسلاش دوبخشی و نرمال آلوده دوبخشی است که اولی دم‌سبک و بقیه دم‌سنگین هستند و ساختار سلسله مراتبی آن به گونه‌ای است که استفاده از الگوریتم امیدگیری– بیشینه‌سازی (EM) در مدل رگرسیون چندکی را میسر می‌سازد. از این خانواده برای مدل‌بندی و برآورد در یک سری داده‌ی واقعی استفاده خواهیم کرد و شاخصی برای برازندگی مناسبت مدل نیز ارائه خواهیم داد.

واژه‌های کلیدی: رگرسیون چندکی، الگوریتم امیدگیری– بیشینه‌سازی، توزیع‌های آمیخته–مقیاس نرمال دوبخشی.

رده‌بندی ریاضی: 33C45; 49-XX

۱ مقدمات

مدل‌های رگرسیون چندکی[۲] (QR) از زمان تحقیق اصلی کوئنکر و باست (۱۹۷۸) و کوئنکر (۲۰۰۵) به طور فزاینده‌ای محبوبیت یافته و برای برازش روی داده‌های واقعی به کار برده شده‌اند. برخلاف مدل رگرسیون میانگین (معمولی)، رگرسیون چندکی به یک خانواده مدل انعطاف پذیرتر تعلق دارد که می تواند ارزیابی کلی از اثرات متغیرهای کمکی در چندک‌های مختلف را ارائه دهد. به طور خاص، ما می‌توانیم چندک‌های پایین‌تر یا بالاتر از پاسخ‌ها را مدل‌سازی کنیم تا یک ارزیابی طبیعی از اثرات متغیرهای کمکی خاص برای آن چندک‌های رگرسیونی ارائه کنیم. بر خلاف مدل‌های مرسوم، که فقط به میانگین شرطی یا اثرات مرکزی متغیرهای کمکی می‌پردازند، مدل‌های رگرسیون چندکی توزیع شرطی متغیر پاسخ را مد نظر قرار می‌دهند، علاوه بر این، رگرسیون چندکی فرضیات توزیعی اضافی را بر خطای مدل تحمیل نمی‌کند.

[۱] نویسنده مسئول مقاله: (M. Maleki) m.maleki.stat@gmail.com
[2] Quantile regression (QR)

از دیدگاه استنباط کلاسیک، بنیتس و همکاران (۲۰۱۳)، ژو و همکاران (۲۰۱۴) و تیان و همکاران (۲۰۱۴) یک مدل رگرسیون چندکی خطی را بر اساس الگوریتم امیدگیری- بیشینه‌سازی[۳] (EM) برآوردگرهای بیشینه درستنمایی (ML) را با فرض لاپلاس نامتقارن[۴] (ALD) بودن توزیع احتمالی خطاها تنظیم کردند. در حالی که ALD دارای خاصیت چندک صفر و یک نمایش تصادفی مفید است، در صفر تشخیص پذیر[۵] نیست، که می‌تواند منجر به مشکلات ناپایداری عددی در برآوردیابی‌ها شود. برای برطرف‌کردن این کمبود، توزیع‌های آمیخته–مقیاس نرمال دوبخشی[۶] (TP-SMN) استفاده شده در قصمی و همکاران (۲۰۲۰) را در ساختار رگرسیون چندکی پیشنهاد خواهیم داد.

خانواده‌ی توزیع‌های احتمالی TP-SMN شامل توزیع‌های احتمالی متقارن- نامتقارن دم سبک- سنگین هستند که رفتار نامتقارن و مشاهدات دورافتاده‌ی موجود در داده‌ها را در ساختار مدل کنترل می‌کنند و یک مدل استوار[۷] برازش می‌دهند. این خانواده تعمیم خانواده‌ی توزیع‌های نرمال مقیاس آمیخته[۸] (SMN) معرفی شده توسط آندریوز و مالوز (۱۹۷۴) به فرم نامتقارن است. نمایش تصادفی این توزیع و ساختار سلسله مراتبی آن شرایط مطلوبی را مهیا می‌سازد تا از آن در ساختار رگرسیون چندکی استفاده کنیم و همچنین قابلیت به کارگیری روش‌های زنجیره مارکوف مونت کارلو (MCMC) و یک الگوریتم EM برای برآوردیابی پارامترهای مدل را فراهم می‌سازد.

ساختار مقاله به شرح زیر است: بخش ۲ شامل معرفی اجمالی از توزیع‌های TP-SMN و ویژگی‌های آن است. در بخش ۳ مدل استوار رگرسیون چندکی برپایه‌ی توزیع‌های TP-SMN را معرفی می‌کنیم و برآوردگرهای بیشینه درستنمایی پارامترهای مدل را با بهره‌گیری از یک الگوریتم EM بدست خواهیم آورد. برازش مدل مزبور بر یک مجموعه داده شامل برخی از ویژگی‌های جسمانی ورزشکاران استرالیا موجود از موسسه ورزش استرالیا (AIS) را در بخش ۴ ارائه می‌دهیم و در بخش ۵ مقاله ضمن نتیجه‌گیری، برخی از مسیرهای تحقیقاتی آینده را نیز ترسیم خواهیم کرد.

۲ خانواده‌ی توزیع‌های آمیخته–مقیاس نرمال دوبخشی

اعضای خانواده‌ی معروف توزیع‌های SMN دارای خاصیت انعطاف‌پذیری بسیار بالایی هستند که هم توزیع‌های دم سبک[۹] و هم دم سنگین[۱۰] را در بردارد. معروف‌ترین اعضای این خانواده توزیع‌های نرمال[۱۱]، تی[۱۲]، اسلاش[۱۳] و نرمال آلوده[۱۴] است، که اولین مورد دم سبک و بقیه دم سنگین هستند. فرم عمومی تابع چگالی اعضای خانواده‌ی SMN به‌صورت زیر است:

$$f_{SMN}\left(y|\mu,\sigma^2,\boldsymbol{\nu}\right)=\int_0^{\infty}\phi\left(y;\mu,u^{-1}\sigma^2\right)dH\left(u;\boldsymbol{\nu}\right),y\in\mathbb{R},$$

که در این چگالی، $\phi(.)$ تابع چگالی نرمال و $H(.)$ معرف تابع توزیع متغیر مقیاس‌ساز مثبت U است. در حقیقت یک آمیخته‌ی نامتناهی از ضریب U^{-1} بر روی واریانس توزیع نرمال اعضای این خانواده را می‌سازد. در نظر گرفتن توزیع‌های متنوع برای متغیر مقیاس ساز U منجر به تولید اعضای خانواده‌ی توزیع‌های آمیخته‌ی مقیاسی نرمال می‌شود. در این خانواده $\mu\in\mathbb{R}$ پارامتر مکان، $\sigma\in\mathbb{R}^+$ پارامتر مقیاس و $\boldsymbol{\nu}$ بردار پارامتر متغیر مقیاس‌ساز است. برخی از کاربردی ترین اعضای خانواده‌ی SMN در جدول ۱ آورده شده‌اند:

جدول ۱: اعضای خانواده‌ی نرمال آمیخته مقیاس

توزیع	$f_{SMN}(.),y\in\mathbb{R}$	توزیع U	$k_r=E\left[U^{\frac{-r}{2}}\right]$
TP–N	$\phi\left(y;\mu,\sigma^2\right)$	$U=1\quad w.p.1$	۱
TP–T	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma}}\left(1+\frac{d}{\nu}\right)^{-\frac{\nu+1}{2}};\boldsymbol{\nu}=\nu$	$Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right)$	$\left(\frac{\nu}{2}\right)^{\frac{r}{2}}\frac{\Gamma\left(\frac{\nu-r}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)},\nu>r$
TP–SL	$\nu\int_0^1 u^{\nu-1}\phi\left(y;\mu,u^{-1}\sigma^2\right)du,\boldsymbol{\nu}=\nu$	$Beta(\nu,1),\nu>0$	$\frac{2\nu}{2\nu-r},\nu>\frac{r}{2}$
TP–CN	$\nu\phi\left(y;\mu,\tau^{-1}\sigma^2\right)+\left(1-\nu\right)\phi\left(y;\mu,\sigma^2\right),\boldsymbol{\nu}=\left(\nu,\tau\right)^{\top}$	$\nu I_{(u_t=\tau)}+\left(1-\nu\right)I_{(u_t=1)},0<\nu<1,0<\tau\leq1$	$\frac{\nu}{\tau^{\frac{r}{2}}}+1-\nu$

تعریف ۱.۲. ساده‌ترین عضو خانواده‌ی توزیع‌های آمیخته- مقیاس نرمال دوبخشی *(TP-SMN)*، توزیع دم سبک نرمال دوبخشی *(TP-N)* با تابع چگالی زیر است:

$$tp.n\left(y;\mu,\sigma,\lambda\right)=\begin{cases}2\lambda\phi\left(y;\mu,\lambda^2\sigma^2\right), & y\leq\mu\\ 2\left(1-\lambda\right)\phi\left(y;\mu,\left(1-\lambda\right)^2\sigma^2\right), & y>\leq\mu\end{cases}\qquad(۱.۲)$$

که در آن $\mu\in\mathbb{R}$ پارامتر مکان، $\sigma\in\mathbb{R}^+$ پارامتر مقیاس و $\lambda\in(0,1)$ پارامتر شکل است. برای تعمیم به چگالی‌های دم سنگین و

[3] Expectation-Maximization (EM-)algorithm
[4] Asymmetric Laplace distribution (ALD)
[5] Identifiable
[6] Two-Piece scale mixtures of normal (TP-SMN)
[7] Robust
[8] Scale mixtures of normal (SMN)
[9] Light-tailed
[10] Heavy-tailed
[11] Normal
[12] Student-t
[13] Slash
[14] Contaminated-Normal

همچنین استفاده از فرم مناسب در ساختار رگرسیون چندکی، فرم چگالی ۱.۲ را می‌توان به‌صورت زیر بازنویسی کرد:

$$tp.n\left(y;\mu,\sigma,\lambda\right)=2\phi\left(\rho_{\lambda}\left(\frac{y-\mu}{\sigma}\right);0;1\right), \tag{۲.۲}$$

که در آن $\rho_{\lambda}(z)=z\left(\lambda-I(z<0)\right)$ تابع زیان بازبین[۱۵] و $I(.)$ تابع نشانگر[۱۶] است.

تعریف ۲.۲. فرم عمومی خانواده تابع چگالی‌های *TP-SMN* به‌صورت زیر تعریف می‌شود:

$$g\left(y,\mu,\sigma,\lambda,\boldsymbol{\nu}\right)=2\int_{0}^{\infty}\phi\left(\rho_{\lambda}\left(\frac{y-\mu}{\frac{\sigma}{\sqrt{u}}}\right);0;1\right)dH\left(u;\boldsymbol{\nu}\right),y\in\mathbb{R} \tag{۳.۲}$$

که در آن $\mu\in\mathbb{R}$ پارامتر مکان، $\sigma\in\mathbb{R}^{+}$ پارامتر مقیاس، $\lambda\in(0,1)$ پارامتر شکل و $\boldsymbol{\nu}$ بردار پارامتر متغیر مقیاس‌ساز U است. خانواده فوق با نماد $Y\sim TP-SMN(\mu,\sigma,\lambda,\boldsymbol{\nu})$ نمایش داده خواهند شد، که اعضای آن شامل تی دوبخشی *(TP-T)*، اسلاش دوبخشی *(TP-SL)* و نرمال آلوده دوبخشی *(TP-CN)* هستند.

گزاره ۳.۲. اگر $Y\sim TP-SMN(0,1,\lambda,\boldsymbol{\nu})$، در اینصورت تابع توزیع متغیر تصادفی استاندارد Y به‌صورت زیر خواهد بود:

$$G(y;0;1,\gamma)=\begin{cases}2\lambda F_{SMN}\left(\frac{y}{1-\gamma};\boldsymbol{\nu}\right), & y\leq 0\\ 2\lambda-1+2(1-\lambda)F_{SMN}\left(\frac{y}{\gamma};\boldsymbol{\nu}\right), & y>0\end{cases}$$

که در آن $F_{SMN}\left(.;\boldsymbol{\nu}\right)$ تابع توزیع متغیر استاندارد نرمال آمیخته مقیاس است.

گزاره ۴.۲. اگر $Y\sim TP-SMN(\mu,\sigma,\lambda,\boldsymbol{\nu})$، آنگاه فرم نمایش تصادفی متغیر تصادفی Y به‌صورت زیر خواهد بود:

$$Y=\mu+(1-\lambda)^{s}\left(-\lambda\right)^{1-s}U^{\frac{-1}{2}}\left|W\right| \tag{۴.۲}$$

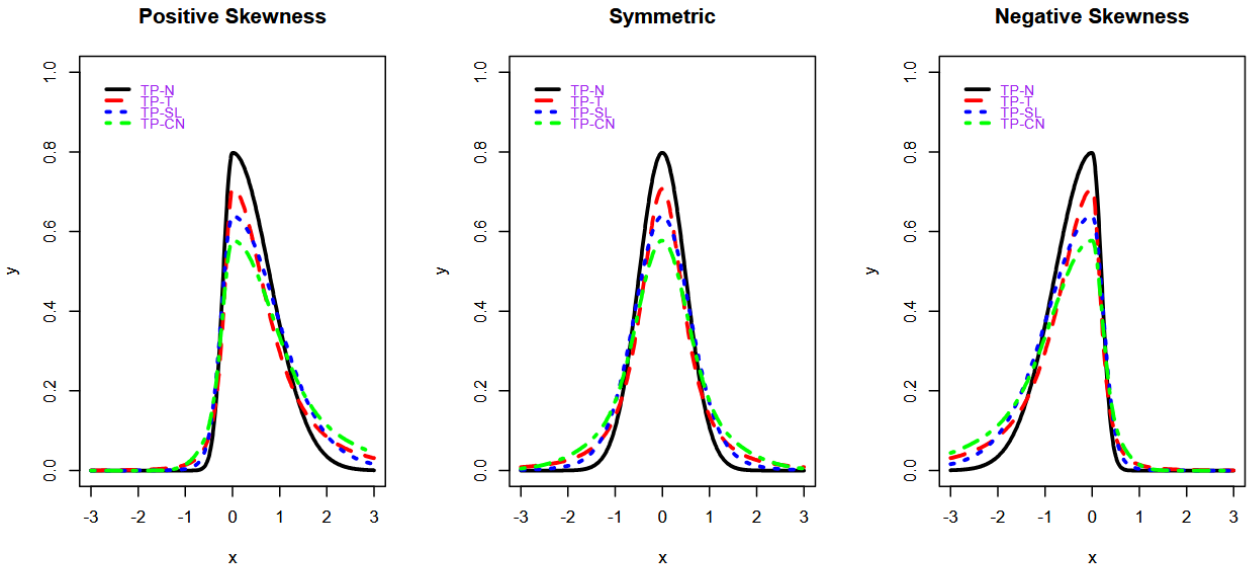
که در آن W ، U و S متغیرهای تصادفی مستقل با توزیع‌های $S\sim Binomial(1,1-\lambda)$ ، $U\sim Gamma(\frac{\nu}{2},\frac{\nu}{2})$ و $W\sim N(0,\sigma^{2})$ هستند. از این فرم در تعیین ساختار سلسله مراتبی مدل استفاده خواهیم کرد. اگر متغیر تصادفی W در نقطه‌ی یک تباهیده (یا اگر $\lambda=0.5$) باشد توزیع‌های نرمال آمیخته‌ی مقیاس حاصل می‌شوند. نمودارهای $TP-SMN$ برای مقادیر چولگی مثبت $(\lambda=0.2)$ متقارن $(\lambda=0.5)$ و چولگی منفی $(\lambda=0.8)$ در حالت استاندارد $\mu=0$ و $\sigma=1$ که برای چگالی $TP-T$ و $TP-SL(\nu=4)$ و $TP-CN(\nu=0.5\tau=1)$ در شکل ۱ رسم شده‌اند.

گزاره ۵.۲. اگر $Y\sim TP-SMN(\mu,\sigma,\lambda,\boldsymbol{\nu})$ باشد، آنگاه

$$E(Y)=\mu+\sigma c_{1}(\gamma)k_{1}d_{1},$$
$$Var(Y)=\sigma^{2}\left[c_{2}(\gamma)k_{2}d_{2}-\left(c_{1}(\gamma)k_{1}d_{1}\right)^{2}\right]$$

که در آن $c_{r}=(1-\lambda)^{r+1}+(-1)^{r}\lambda^{r+1}$، $k_{r}=E\left[U^{\frac{-r}{2}}\right]$ (تعریف شده در جدول ۱) در، $d_{2m}=\frac{(2m)!}{2^{m}m}$ و $d_{2m+1}=m!2^{m}\sqrt{\frac{2}{\pi}}$ جزئیات بیشتر در مورد خانواده‌ی $TP-SMN$ را در ملکی و همکاران (۲۰۱۹) و بذرافکن و همکاران (۲۰۲۲) و زارعی و همکاران (۲۰۲۳) ببینید.

[15] Check function
[16] Indicator



شکل ۱: نمودار چگالی‌های آمیخته–مقیاس نرمال دوبخشی

۳ رگرسیون چندکی برپایه‌ی خانواده‌ی توزیع‌های آمیخته–مقیاس نرمال دوبخشی

فرض کنید $Y_i, \quad i = 1, \cdots, n$ متغیر پاسخ و $\boldsymbol{x_i} = (x_{i1}, \cdots, x_{ip})^\top$ بردار متغیرهای کمکی $p \times 1$ بعدی برای مشاهده iام و $Q_{Y_i}(\lambda|\boldsymbol{x_i})$ چندک شرطی λ–ام $Y_i(0 < \lambda < 1)$ نسبت به $\boldsymbol{x_i}$ باشد، یعنی $Q_{Y_i}(\lambda \boldsymbol{x_i}) = inf(\{y_i|F_{y_i|\boldsymbol{x_i}})(y_i) \leq \lambda\} =$ $F^{-1}_{y_i|\boldsymbol{x_i}}(\lambda)$. در رگرسیون چندکی فرض می‌شود رابطه‌ی بین چندک شرطی λ–ام و $\boldsymbol{x_i}$ یعنی $Q_{Y_i}(\lambda|\boldsymbol{x_i})$ به فرم $Q_{Y_i}(\lambda|\boldsymbol{x_i}) = \boldsymbol{x_i}^\top \boldsymbol{\beta}_\lambda$ مدل‌بندی شود که در آن $\boldsymbol{\beta}_\lambda$ یک بردار $p \times 1$ بعدی باشد. لذا در این حالت مدل رگرسیون چندکی را به‌صورت رابطه‌ی زیر در نظر می‌گیریم:

$$Q_{Y_i}(\lambda|\boldsymbol{x_i}) = \boldsymbol{x_i}^\top \boldsymbol{\beta}_\lambda + \varepsilon_i, \quad i = 1, \cdots, n, \tag{۱.۳}$$

که در آن ε_i خطای مدل است و توزیع احتمالی آن به‌گونه‌ای تعریف می‌شود که چندک λ–ام آن صفر باشد، که در نتیجه خواهیم داشت $P(Y_i \leq \boldsymbol{x_i}^\top \boldsymbol{\beta}_\lambda) = \lambda$. با در نظر گرفتن $\lambda = 0.5$ رگرسیون میانه‌ای را خواهیم داشت و در حالت کلی بردار پارامتر ضرایب رگرسیون $\boldsymbol{\beta}_\lambda$ با رابطه‌ی زیر برآورد می‌شود.

$$\hat{\boldsymbol{\beta}}_\lambda = argmin_{\boldsymbol{\beta}_\lambda \in \mathbb{R}^p} \left(\sum_{i=1}^{n} \rho_\lambda \left(Y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}_\lambda \right) \right) \tag{۲.۳}$$

که در آن $\rho_\lambda(.)$ تابع بازبین و $\hat{\boldsymbol{\beta}}$ حاصله برآورد $\boldsymbol{\beta}$ در چندک λ –ام توزیع شرطی Y به شرط $\boldsymbol{x_i}$ است. باتوجه به ساختار چگالی TP-SMN در ۳.۲ کمینه ساختن در ۲.۳ معادل بیشینه ساختن تابع درستنمایی مدل خواهد بود، لذا با در نظر گرفتن ساختار سلسله مراتبی خانواده TP-SMN در بخش بعد از یک الگوریتم EM برای برآوردیابی پارامترهای مدل استفاده خواهیم نمود.

۱.۳ برآورد پارامترهای مدل با الگوریتم امیدگیری بیشینه‌سازی:

الگوریتم عددی ECME یک تعمیم از الگوریتم ECM است که توسط منگ و روبین (۱۹۹۳) معرفی شد، که به نوبه خود تعمیمی از الگوریتم عددی EM است که توسط دمپستر و همکاران (۱۹۷۷) معرفی گردید. در الگوریتم مورد استفاده ECME در آخرین مرحله از قسمت

ماکزیمم‌سازی شرطی، تابع لگاریتم درستنمایی اصلی را برای برخی از پارامترها ماکزیمم می‌کنند. در این قسمت برای برآوردیابی پارامترهای مدل رگرسیون چندکی با کمک متغیرهای پنهان از این الگوریتم استفاده خواهیم کرد. باتوجه به چگالی‌های دم‌سنگین TP-SMN در ۳.۲ و بر پایه چگالی دم‌سبک TP-N در ۲.۲، فرم سلسله مراتبی زیر را خواهیم داشت:

$$Y_i|U_i = u_i \sim TP - N\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_\lambda, \frac{\sigma}{u_i}\right),$$
$$U_i \sim H(u_i; \boldsymbol{\nu}).$$

اگر مقادیر مشاهده‌شده‌ی متغیر پاسخ را با $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$ و مقادیر متغیر پنهان را با $\boldsymbol{U} = (U_1, \cdots, U_n)^\top$ نمایش دهیم، براساس فرم سلسله مراتبی فوق تابع درستنمایی برمبنای مشاهدات کامل و بردار پارامتر $\boldsymbol{\Theta} = \left(\boldsymbol{\beta}_\lambda^\top, \sigma, \boldsymbol{\nu}\right)^\top$ به‌صورت زیر خواهد بود:

$$l_c(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{u}) = \sum_{i=1}^{n} \log \phi(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}_\lambda, \sigma u_i^{-1} \eta_i^{-1}) + \sum_{i=1}^{n} \log h(u_i; \boldsymbol{\nu}),$$

که در آن $\eta_i = \lambda I\left(y_i \leq \boldsymbol{x}_i^\top \boldsymbol{\beta}_\lambda\right) + (1-\lambda) I\left(y_i > \boldsymbol{x}_i^\top \boldsymbol{\beta}_\lambda\right)$ است. اگر بروزرسانی پارامترها در مرحله r ام را با $\boldsymbol{\Theta}^{(r)}$ نمایش دهیم، در مرحله امیدگیری از الگوریتم تابع $Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(r)}\right) = E\left(l_c(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{u})|\boldsymbol{y}, \boldsymbol{\Theta}^{(r)}\right)$ داریم:

$$Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(r)}\right) = c - n\log\sigma - 2\sum_{i=1}^{n} \hat{u}_i \eta_i^2 \frac{\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\lambda\right)^2}{\sigma^2} + \sum_{i=1}^{n} E\left(\log h(u_i; \boldsymbol{\nu})|\boldsymbol{y}, \boldsymbol{\Theta}^{(r)}\right)$$

که مقدار ثابت c حاوی پارامتر نیست و برای مشخص شدن ساختار این تابع برای مرحله‌ی بیشینه‌سازی لازم است مقادیر $\hat{u}_i = E(u_i|\boldsymbol{y}, \boldsymbol{\Theta}^{(r)}))$ را برای اعضای متناظر خانواده‌ی TP-SMN به‌صورت زیر مشخص نمود:

الف) نرمال دوبخشی (TP-N):

$$\hat{u}_i = 1, \quad i = 1, \cdots, n,$$

ب) تی دوبخشی (TP-T):

$$\hat{u}_i = \frac{\nu + 1}{\nu + d_i}, \quad i = 1, \cdots, n,$$

ج) اسلاش دوبخشی (TP-SL):

$$\hat{u}_i = \frac{2\nu + 1}{d} \frac{p_{1\left(\frac{\nu+3}{2}, \frac{d_i}{2}\right)}}{p_{1\left(\frac{\nu+1}{2}, \frac{d_i}{2}\right)}},$$

د) نرمال آلوده دوبخشی (TP-CN):

$$\hat{u}_i = \frac{\tau^2 \nu e^{-\frac{\tau d_i}{2}} + (1-\nu) e^{-\frac{d_i}{2}}}{\tau \nu e^{-\frac{\tau d_i}{2}} + (1-\nu) e^{-\frac{d_i}{2}}},$$

که در آنها $d_i = \left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\lambda}{\sigma^2}\right)^2$ و $P_x(a,b)$ نمایش دهنده تابع توزیع تجمعی $Gamma(a,b)$ است که در نقطه x محاسبه گردیده است.

در مرحله‌ی بیشینه‌سازی، بروزرسانی پارامترها در هرچرخه از الگوریتم به‌صورت زیر خواهند بود:

$$\hat{\boldsymbol{\beta}}_\lambda^{(r+1)} = \left(\sum_{i=1}^{n} \hat{u}_i \eta_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \sum_{i=1}^{n} \hat{u}_i \eta_i^2 y_i \boldsymbol{x}_i,$$
$$\hat{\sigma^2}^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i \eta_i^2 \left(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda^{(r+1)}\right)^2,$$

و در مرحله آخر الگوریتم ECME بروزرسانی پارامتر $\boldsymbol{\nu}$ به‌صورت زیر خواهد بود:

$$\boldsymbol{\upsilon}^{(k+1)} = argmax_{\boldsymbol{\nu}} \sum_{i=1}^{n} \log f\left(y_i; \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(r+1)}, \hat{\sigma}^{(r+1)}, \boldsymbol{\nu}\right)$$

و الگوریتم ECME تا زمانی که $\left|\frac{l(\hat{\boldsymbol{\Theta}}^{(r+1)})}{l(\hat{\boldsymbol{\Theta}}^{(r)})} - 1\right| \leq 10^{-4}$ باشد ادامه می‌یابد.

۴ کاربرد رگرسیون چندکی TP-SMN روی ویژگی های جسمانی ورزشکاران استرالیا (AIS)

در این بخش مدل رگرسیون چندکی مفروض را بر روی مجموعه داده‌های موجود در کوک و ویزبرگ (۱۹۹۴) در مورد ویژگی‌های جسمانی ورزشکاران استرالیا موجود در موسسه ورزش استرالیا (AIS) ، برازش می‌دهیم. برای این منظور از متغیر شاخص توده بدنی (bmi) به عنوان متغیر پاسخ، و از متغیرهای توده بدن بدون چربی (lbm) و جنسیت (sex) به عنوان متغیرهای پیش‌بینی کننده‌ی مرتبط با 202 ورزشکار استرالیایی استفاده می‌کنیم. برای اجرای الگوریتم ECME از نرم افزار R نسخه 4.4.2 با سخت افزار پردازنده $corei7760$ با فرکانس 2.8 گیگاهرتز استفاده کردیم. برای مقادیر اولیه در الگوریتم، ابتدا با استفاده از روش حداقل مربعات برآورد ضرایب رگرسیونی را به‌دست آوردیم، سپس با محاسبه اولیه برآوردهای خطای مدل، با استفاده از روش گشتاوری مقادیر اولیه پارامترهای خانواده آمیخته–مقیاس نرمال دوبخشی را محاسبه کردیم. پیچیدگی محاسباتی الگوریتم ارائه شده به‌صورت $O(nkt)$ می‌باشد که در آن n حجم نمونه، k تعداد پارامترهای مدل و t تعداد تکرارهای لازم برای رسیدن به همگرایی می باشد. این الگوریتم از لحاظ محاسباتی کارآمد بوده و در مقایسه با روش‌های مشابه، همگرایی مطلوبی دارد. الگوریتم پیشنهادی بر روی یک سیستم محاسباتی استاندارد، با آستانه تحمل 0.001 و پس از 30 تکرار، در مدت زمان 25 ثانیه به همگرایی رسید (معادل میانگین 0.83 ثانیه برای هر تکرار). در پایان نکات لازم برای ارائه محاسبات عددی، توجه داشته باشید که بهینه‌سازی‌های انجام شده در پیاده‌سازی الگوریتم موجب کاهش قابل توجه هزینه‌های محاسباتی شده‌اند.

مدل رگرسیون خطی کلاسیک عبارت است از:

$$bmi_i = \beta_0 + \beta_1 lbm_i + \beta_2 sex_i + \varepsilon_i, \quad i = 1, \cdots, 202,$$

که درآن خطای مدل از توزیع احتمالی خانواده‌ی TP-SMN تبعیت می‌کند. با در نظر گرفتن رگرسیون خطی فوق در دو حالتی که خطاها دارای توزیع متقارن و نامتقارن باشند، مقادیر معیارهای انتخاب مدل به شرح جدول ۲ خواهد بود. حال با درنظر گرفتن رگرسیون میانه‌ای

جدول ۲: معیارهای مناسبت مدل برای مقایسه مدل‌های رگرسیون خطی متقارن و نامتقارن.

معیار	نرمال	نرمال دوبخشی	تی	تی دوبخشی	اسلاش	اسلاش دوبخشی	نرمال آلوده	نرمال آلوده دوبخشی
AIC	1232.88	1177.30	942.64	895.44	913.00	889.66	929.84	912.82
BIC	1264.11	1190.53	959.18	911.98	929.54	906.20	949.68	932.66
Log-likelihood	−612.44	−584.65	−466.32	−442.72	−451.50	−439.83	−458.92	−450.41

$(\lambda = 0.5)$ و بر اساس معیارهای مناسبت مدل آکائیکی (AIC، 1974)، شوارتز (BIC، 1978) و تابع درستنمایی بیشینه شده (Log-likelihood)، برای مقایسه بین مدل‌ها، مناسبترین برازش را انتخاب خواهیم کرد. جدول ۳ نتایج به دست آمده را برای هر معیار مقایسه مدل ارائه می‌کند. با مقایسه دو جدول فوق ملاحظه می‌شود که مدل رگرسیون چندکی به ازای $\lambda = 0.5$ نه تنها برازش بهتری نسبت

جدول ۳: معیارهای مناسبت مدل برای مقایسه مدل‌های رگرسیون‌چندکی میانه‌ای متقارن و نامتقارن.

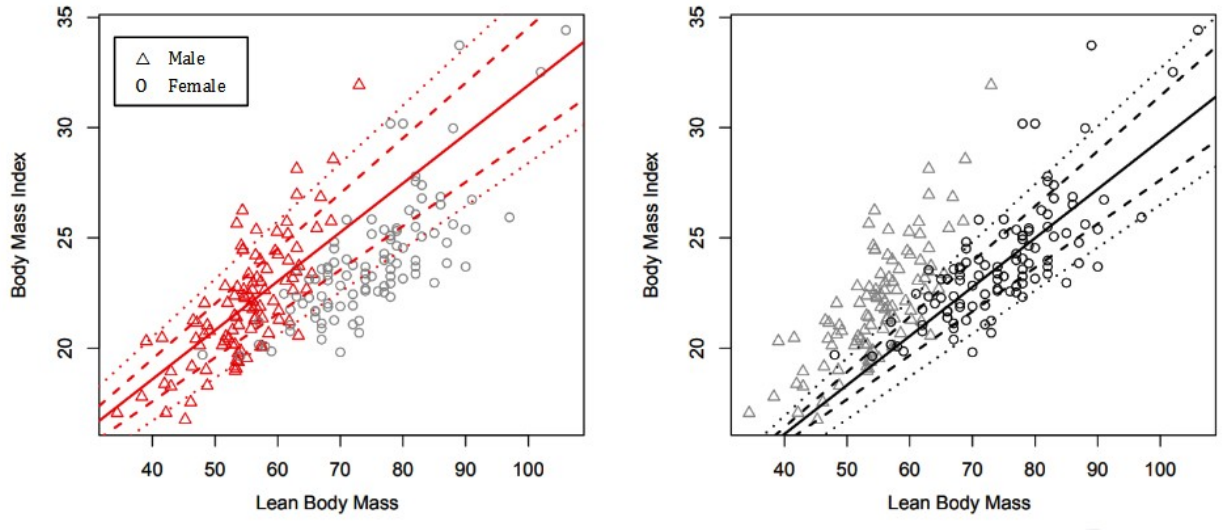
معیار	نرمال دوبخشی	تی دوبخشی	اسلاش دوبخشی	نرمال آلوده دوبخشی
AIC	1062.62	802.40	804.60	810.22
BIC	1075.85	818.94	821.14	830.07
Log-likelihood	−527.31	−396.20	−397.30	−399.11

به رگرسیون خطی دارد بلکه با انعطاف‌پذیری بالای مدل چندکی، می‌توان برای چندک‌های مختلف توزیع شرطی متغیر پاسخ به کار برد

جدول ۴: برآورد و انحراف معیار پارامترهای مدل رگرسیون میانه‌ای تی دوبخشی.

پارامترها	β_0	β_1	β_2	σ	ν
برآورد	7.32	0.25	2.61	1.13	2.04
انحراف معیار	1.70	0.05	0.48	0.15	0.03

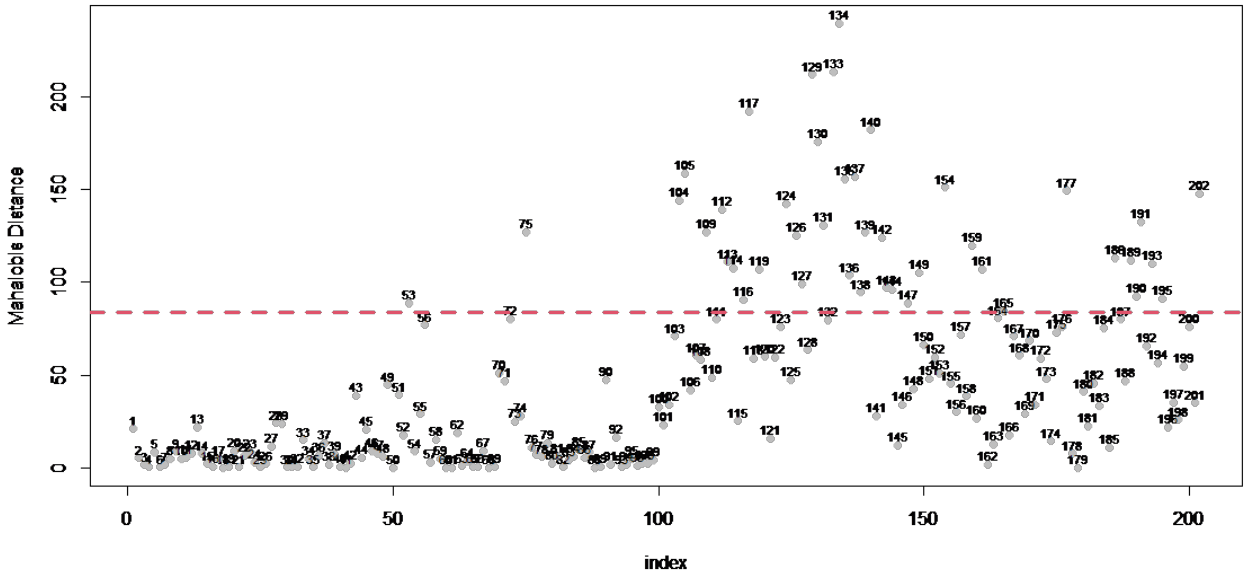
که استواری بیشتری در مقابل مشاهدات پرت خواهد داشت. با توجه به مقادیر معیارهای جدول ۳ ، می توان نتیجه گرفت که مدل استوار رگرسیون میانه‌ای تی دوبخشی (و تقریبا اسلاش دوبخشی) بهتر از سایر مدل‌ها است. رگرسیون‌های چندکی متفاوتی (باλ های متفاوت شامل مقادیر0.15، 0.30، 0.50، 0.70،0.85) از این توزیع را در شکل ۲ ملاحظه می‌کنید. همچنین می‌توان استنباط کرد که مدل‌های رگرسیون دمسنگین بهتر از مدل دمسبک عمل می‌کنند که دلیل آن وجود مشاهدات پرت است.



شکل ۲: رگرسیون چندکی تی دوبخشی برازش شده شاخص توده‌ی بدنی روی جنسیت (مردان سمت چپ نقاط با رنگ قرمز پررنگ و زنان سمت راست نقاط با رنگ مشکی پررنگ،) با چندک‌های 0.15، 0.30، 0.50، 0.70،0.85 روی داده‌های AIS.

برآورد پارامترهای مدل رگرسیون میانه‌ای تی دوبخشی با انحراف معیار برآوردها که با روش بوت استرپ بدست آمده‌اند نیز در جدول ۴ گزارش شده‌اند. در مدل رگرسیون چندکی تی دوبخشی چنانچه برآورد پارامتر درجه آزادی (ν) مقدار بالایی داشته باشد توزیع دمسنگین تی دوبخشی معادل توزیع دمسبک نرمال دوبخشی خواهد بود که نشان از عدم وجود مشاهدات پرت می‌باشد، و چنانچه برآورد پارامتر درجه آزادی مانند مقدار 2.04 در داده‌های فوق مقدار کوچکی باشد نشان از وجود مشاهده پرت دارد. لذا این مدل در صورت وجود یا عدم وجود مشاهدات پرت به صورت هوشمند عمل می‌کند، لذا ضرورتی به شناسایی مشاهدات پرت، قبل از برازش مدل نیست. با توجه به پیشنهاد کاربردی یکی از داوران مقاله، ما به شناسایی مقادیر پرت در این مثال نیز خواهیم پرداخت.

در رگرسیون میانه‌ای آمیخته- مقیاس نرمال دوبخشی با در نظر گرفتن فرم نمایش تصادفی متغیر تی دوبخشی و مدل رگرسیون چندکی، فاصله ماهالانوبیس برای مدل برازش شده به‌صورت $\triangle_i = \frac{4(y_i - \mathbf{x}_i^\top \beta_\lambda)^2}{\sigma^2}, i = 1, \cdots, 202$ برای اعضای مختلف خانواده آمیخته- مقیاس نرمال دوبخشی شامل توزیع نرمال دوبخشی به‌صورت $\triangle_i \sim \chi_1^2$، توزیع تی دوبخشی به‌صورت $\triangle_i \sim F(1, \nu)$، توزیع اسلاش دوبخشی به‌صورت $Pr(\triangle_i \leq s) = Pr(\chi_1^2 \leq s) - \frac{2^\nu \Gamma\left(\frac{\nu+1}{2}\right)}{s^\nu \Gamma\left(\frac{1}{2}\right)} Pr(\chi_{2\nu+1}^2 \leq s)$ و توزیع نرمال آلوده‌ی دوبخشی به‌صورت $Pr(\triangle_i \leq s) = \nu Pr(\chi_1 1^2 \leq \gamma s) + (1-\nu) Pr(\chi_1^2 \leq s)$ خواهد بود، در این راستا زلر و همکاران (۲۰۱۰) را نیز ببینید، لذا با در نظر گرفتن چندک 0.99 (خط نقطه چین در شکل۳) از توزیع منتخب برازش شده‌ی تی دوبخشی، مشاهدات پرت در بالای این مقدار در شکل ۳ قرار دارند.



شکل ۳: فاصله ماهالانوبیس برای مدل برازش شده رگرسیون میانه‌ای تی دوبخشی روی داده‌های AIS.

در مرحله‌ی آخر برای تایید نیکویی برازش مدل مزبور، با محاسبه باقیمانده‌های مدل به‌صورت $e_i = y_i - \mathbf{x}_i^\top \beta_\lambda, i = 1, \cdots, 202$ محاسبه شده‌اند که انتظار می‌رود از توزیع تی دوبخشی با پارامترهای برآورد شده تبعیت کنند. آماره کولموگروف–اسمیرنوف (K–S) مقدار جزئی $D = 0.043$ را پذیرفت و همچنین با انجام روش بوت استرپ، 500 مرتبه داده‌ها را برای مدل شبیه سازی کردیم و پی–مقدار $p \approx 0.9132$ به‌دست آمد، که این مقادیر نشان از برازش مدل مناسب برپایه توزیع تی دوبخشی دارند.

۵ نتیجه گیری

در این مقاله، یک رویکرد کلاسیک برآوردیابی روی یک مدل استوار رگرسیون چندکی برپایه‌ی توزیع‌های آمیخته–مقیاس نرمال دوبخشی در نظر گرفته شده که رفتار نامتقارن و اثر مشاهدات دورافتاده را نیز کنترل می‌کند. ساختار سلسله مراتبی این خانواده امکان دسترسی به یک الگوریتم امیدگیری– بیشینه‌سازی را برای برآورد پارامترهای مدل فراهم می‌کند. همچنین نتایج عملکرد مناسبت این مدل و روش برآوردیابی روی یک سری داده‌های جسمانی ورزشکاران استرالیا ارائه گردیده است. برای تحقیقات آینده می‌توان مدل‌بندی رگرسیون چندکی را برای داده‌های تحت سانسور، برش داده شده، داده‌های $\alpha-$ پایدار، که توزیع‌های آمیخته–مقیاس می‌تواند برازش مناسبی بر آنها داشته باشد، مورد بررسی و آزمون قرار داد.

سپاس‌گزاری: ازهیئت تحریریه و داوران محترم که بادقت و بررسی‌های موشکافانه‌شان، به بهبود این مقاله کمک کردند، صمیمانه سپاس‌گزاریم.

فهرست منابع

[1] Akaike, H.,1974. A new look at the statistical model identification. IEEE Trans. Autom. Cont., 19, pp.379–397. doi: 10.1109/TAC.1974.1100705

[2] Andrews, D. F. and Mallows, C. L., 1974. Scale mixtures of normal distributions. J. R. Stat. Soc. Ser. B Methodol., 36, pp.99–102.

[3] Bazrafkan, M., Zare, K., Maleki, M. and Khodadadi, Z., 2022. Partially linear models based on heavy-tailed and asymmetrical distributions. Stoch. Environ. Res. Risk. Assess., 36, pp.1243–1253.doi: 10.1007/s00477-021-02101-1

[4] Benites, L., Lachos, V. H. and Vilca, F., 2013. Likelihood based inference for quantile regression using the asymmetric Laplace distribution. Technical Report 15, Universidade Estadual de Campinas.

[5] Cook, R. D. and Weisberg, S., 1994. An Introduction to Regression Graphics. Wiley, New York. doi: 10.1002/9781118150658

[6] Dempster, A., Laird, N. and Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Seri. B, 39, pp.1–38. doi: 10.1111/j.2517-6161.1977.tb01600.x

[7] Ghasami, S., Maleki, M., Khodadadi, Z., 2020. Leptokurtic and platykurtic class of robust symmetrical and asymmetrical time series models, J. Comput. Appl. Math, 376, 112806. doi: 10.1016/j.cam.2020.112806

[8] Koenker, R., 2005. Quantile Regression, volume 38. Cambridge University Press. doi: 10.1017/CBO9780511754098

[9] Koenker, R. & G Bassett, J., 1978. Regression quantiles. Econometrica: J. Econom., 46, pp.33–50. doi: 10.2307/1913643

[10] Maleki, M., Barkhordar, Z., Khodadadi, Z. and Wraith, D.,2019. A robust class of homoscedastic nonlinear regression models. J. Stat. Comput. Simul., 89(14), pp.2765–2781. doi: 10.1080/00949655.2019.1635598

[11] Meng, X. and Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika, 80, pp.267–278. doi: 10.1093/biomet/80.2.267

[12] Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat., 6(2), pp.461–464. doi: 10.1214/aos/1176344136

[13] Tian, Y., Tian, M. and Zhu, Q., 2014. Linear Quantile Regression Based on EM Algorithm. Commun. Stat. - Theory Methods., 43(16), pp.3464–34. doi: 10.1080/03610926.2013.766339

[14] Zarei, A., Khodadadi, Z., Maleki, M.,and Zare, K., 2023. Robust mixture regression modeling based on two-piece scale mixtures of normal distributions. Advances in Data Analysis and Classification, 17, pp.181–210. https://doi.org/10.1007/s11634-022-00495-6

[15] Zeller, C. B., Labra, F. V., Lachos, V. H., Balakrishnan, N., 2010. Influence analyses of skew-normal-independent linear mixed models. Comput. Stat. Data Anal, 54, 1266–1280.

[16] Zhou, Y.-h., Ni, Z.-x. and Li, Y., 2014. Quantile Regression via the EM Algorithm. Commun. Stat. Simul. Comput., 43(10), pp.2162–2172. doi: 10.1080/03610918.2012.746980

Quantile regression based on the two-piece scale mixtures of normal distributions

Mohsen Maleki[(1)] [17] and Alireza Nematollahi[(2)]

[(1)] Department of Statistics, Faculty of Mathematics and Statistics, University of Isfahan, 81746-73441 Isfahan, Iran
[(2)] Department of Statistics, Shiraz University, Shiraz, Iran

Communicated by: Hamzeh Torabi

Received: 2 April 2025 Accepted: 15 November 2025

Abstract: Standard regression models frequently exhibit poor performanc when modeling data with outliers and their asymmetric behavior. Therefore, using symmetric-asymmetric light-heavy tail models in their structure can overcome the above problems. In this article, we will introduce a robust quantile regression model with distributed errors based on two-piece scale mixture of normal (Malaki et al. 2019) to solve the above problems and obtain the model parameter estimates using the maximum likelihood approach. The two-piece scale mixture of normal family includes two-piece normal, two-piece t, two-piece slash, and two-piece contaminated normal distributions, the first of which is light-tailed and the rest are heavy-tailed.The family's hierarchical structure facilitates implementation of the expectation-maximization (EM) algorithm for parameters estimation in the quantile regression frame works. This family for will be used modeling and estimation in a real data set and an index of model goodness of fit will be also provided.

Keywords: EM-algorithm, Quantile regression, Two-piece scale mixture of normal family.

©2024 Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-NonComertial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

[17]Corresponding author :(M. Maleki) m.maleki.stat@gmail.com