مجله علوم آماری، پاییز و زمستان ۱۳۹۶
جلد ۱۱، شماره ۲، ص ۲۹۷ – ۳۱۷
DOI: 10.18869/acadpub.jss.11.2.297

تحلیل بیزی رگرسیون چوله‌نرمال آمیخته

افشین فلاح ، رامین کاظمی، حسن خسروی

گروه آمار، دانشگاه بین المللی امام خمینی

تاریخ دریافت: ۱۳۹۵/۸/۱۳ تاریخ آخرین بازنگری: ۱۳۹۶/۱۱/۲۱

چکیده: تحلیل رگرسیونی به‌طور سنتی با فرض همگن بودن جامعه و نرمال بودن توزیع متغیر پاسخ صورت می‌پذیرد. این در حالی است که در بسیاری از کاربردها، به‌دلیل ناهمگنی مشاهدات، وجود نقاط دور افتاده، چولگی یا ترکیبی از آن‌ها، مشاهدات ساختاری ناهمگن با زیرجوامعی چوله–متقارن را نشان می‌دهند. در چنین حالاتی، می‌توان آمیخته‌ای متناهی از توزیع‌های چوله–متقارن را برای مدل‌بندی جامعه مورد استفاده قرار داد. در این مقاله رهیافت بیزی تحلیل رگرسیونی تحت فرض ناهمگن بودن جامعه و چوله–متقارن بودن توزیع زیرجوامع، با استفاده از آمیخته‌ای متناهی از توزیع‌های چوله‌نرمال مورد توجه قرار گرفته است. به منظور ارزیابی رهیافت پیشنهادی و مقایسه آن با مدل فراوانی‌گرا، از یک مطالعه شبیه‌سازی و یک مثال کاربردی استفاده شده است.

واژه‌های کلیدی: تحلیل رگرسیون بیزی، توزیع آمیخته متناهی، چولگی، الگوریتم EM، الگوریتم گیبز.

۱ مقدمه

نظریه سنتی تحلیل رگرسیونی بر پایه همگن بودن جامعه و نرمال بودن توزیع مشاهدات بنا شده 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. جایگزین کردن توزیع نرمال با توزیع‌های مناسب‌تری که قابلیت مدل‌بندی داده‌های نامتقارن را نیز داشته باشند، راهکاری است که در سال‌های اخیر مورد توجه بسیاری از محققین قرار گرفته است. در این زمینه، می‌توان به مطالعات

آدرس الکترونیکی نویسنده مسئول مقاله: افشین فلاح، a.fallah@sci.ikiu.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J02، 62F15

آرلانوواله و همکاران (۲۰۰۸)، لاچوس و همکاران (۲۰۱۰)، کانچو و همکاران ($۲۰۱۰a$) اشاره کرد، که از توزیع چوله‌نرمال برای برازش مدل رگرسیونی استفاده کرده‌اند. اهمیت توزیع چوله‌نرمال و مبنا قرار گرفتن آن در بسیاری از مباحث استنباط آماری و مدل‌سازی به این دلیل است که این توزیع علی‌رغم داشتن قابلیت مدل‌بندی مشاهدات نامتقارن، شباهت زیادی به توزیع نرمال داشته و آن را نه به‌عنوان یک توزیع حدی، بلکه به‌عنوان یک عضو در بردارد. از طرفی، هنگامی که جامعه ساختاری ناهمگن دارد و از زیرجوامعی چوله-متقارن تشکیل شده است، می‌توان آمیخته‌ای متناهی از توزیع‌های چوله-متقارن را مدنظر قرار داد. توزیع‌های آمیخته متناهی، به دلیل انعطاف‌پذیری قابل توجه آن‌ها، در ادبیات آماری به دفعات در زمینه‌های مختلف مورد استفاده قرار گرفته‌اند و اغلب قادرند توصیف مناسبی از جوامع پیچیده و ناهمگن فراهم سازند. تحلیل رگرسیونی تحت توزیع‌های آمیخته متناهی، اولین بار توسط کوانت (۱۹۷۲) و کوانت و رامسِی (۱۹۷۸) مورد مطالعه قرار گرفت. با معرفی الگوریتم EM توسط دمپستر و همکاران (۱۹۷۷)، محققانی بسیاری از جمله اِتکین و ویلسون (۱۹۸۰)، دساربو و کورن (۱۹۸۸)، تورنر (۲۰۰۰)، لاچوس و همکاران (۲۰۰۱، ۲۰۱۰)، کانچو و همکاران ($۲۰۱۰b$) و لیو و همکاران (۲۰۱۱) تحلیل رگرسیونی چوله‌نرمال را از دیدگاه فراوانی‌گرایانه مورد مطالعه قرار دادند. با توجه به دشواری‌های ذاتی برآورد پارامترها در توزیع‌های آمیخته متناهی، و این که خواص بهینه برآوردگرهای ماکسیمم درستنمایی تنها در حالت مجانبی تحقق می‌یابند، رهیافت فراوانی‌گرایانه تحلیل رگرسیونی تحت توزیع‌های آمیخته متناهی از کارایی لازم برخوردار نیست. نشان داده شده است تابع درستنمایی مبتنی بر توزیع چوله‌نرمال، به خصوص در حالت آمیخته بودن توزیع، نسبت به پارامترهای چولگی کران‌دار نیست. افزون بر این، در زمینه کاربست الگوریتم EM به‌عنوان عمومی‌ترین راه برای محاسبه برآوردگرهای ماکسیمم درستنمایی در توزیع‌های آمیخته متناهی، ملاحظات و دشواری‌هایی وجود دارد، که از آن جمله می‌توان به عدم همگرایی الگوریتم EM در برآورد پارامترهای چولگی اشاره کرد. این مشکلات توسط برناردی و همکاران (۲۰۱۲) مورد توجه قرار گرفته است. با توجه به مشکلات رهیافت فراوانی‌گرا، در این مقاله رهیافت بیزی تحلیل رگرسیونی آمیخته متناهی چوله‌نرمال مورد مطالعه قرار گرفته است. برای این منظور، با در نظر گرفتن توزیع‌های پیشین مناسب، توزیع‌های پسین شرطی کامل پارامترهای مدل محاسبه و یک الگوریتم گیبز برای نمونه‌گیری از توزیع پسین توسعه داده شده است.

در بخش ۲، رهیافت فراوانی‌گرایانه تحلیل رگرسیونی آمیخته متناهی چوله‌نرمال، به صورت اجمالی مورد بحث قرار می‌گیرد. برای این منظور، توزیع چوله‌نرمال و برخی از ویژگی‌های شاخص آن با تاکید بر مشکلات مربوط به برآورد پارامترهای این توزیع در رهیافت فراوانی‌گرا به اختصار شرح داده می‌شود. در

بخش ۳، برای تحلیل رگرسیونی تحت مدل چوله‌نرمال آمیخته متناهی یک مدل بیزی پیشنهاد شده است. برای ارزیابی مدل بیزی پیشنهادی و مقایسه آن با همتای فراوانی‌گرایانه رقیب، یک مطالعه شبیه‌سازی در بخش ۴ اجرا شده است. در بخش ۵، یک مثال کاربردی در زمینه هزینه‌های خالص خانوارهای شهری در ایران استفاده شده است. بحث و نتیجه‌گیری در بخش ۶ آورده شده است.

۲ تحلیل فراوانی‌گرایانه رگرسیون چوله‌نرمال آمیخته

متغیر تصادفی Z دارای توزیع چوله‌نرمال با پارامتر چولگی α، پارامتر مکان ξ و پارامتر مقیاس ω است ($SN(\xi,\omega,\alpha)$)، هرگاه تابع چگالی آن به‌صورت

$$f_{SN}(y|\xi,\omega,\alpha)=\frac{۲}{\omega}\phi(\frac{y-\xi}{\omega})\Phi(\alpha\frac{y-\xi}{\omega}),\qquad \omega>۰,\xi,\alpha\in\mathbb{R}, \tag{۱}$$

باشد، که در آن $\phi(\cdot)$ و $\Phi(\cdot)$ به‌ترتیب نشان دهنده توابع چگالی و توزیع تجمعی متغیر تصادفی نرمال استاندارد هستند (آزالینی، ۱۹۸۵). این توزیع در واقع آمیخته‌ای از دو توزیع نرمال و نیم‌نرمال است. فرض کنید متغیرهای تصادفی مستقل U و T به ترتیب دارای توزیع‌های نرمال $N(۰,\omega^۲)$ و نرمال بریده‌شده $TN(۰,\omega^۲)\mathbb{I}_{(t)}(۰,+\infty)$ باشند. در این‌صورت متغیر تصادفی

$$Y=\xi+\frac{\alpha}{\sqrt{۱+\alpha^۲}}T+\frac{۱}{\sqrt{۱+\alpha^۲}}U, \tag{۲}$$

دارای توزیع (۱) خواهد بود و

$$T|Y=y\sim N(\delta(\alpha)(y-\xi),\omega^۲(۱-\delta^۲(\alpha)))\mathbb{I}_{(t)}(۰,+\infty). \tag{۳}$$

به‌علاوه، به کمک روابط مربوط به گشتاورهای توزیع نرمال بریده‌شده (به‌عنوان مثال، بار و همکاران، ۱۹۹۹ را ببینید)، می‌توان نشان داد که دو گشتاور اول توزیع (۳) که در اجرای الگوریتم EM مورد نیاز هستند، به‌ترتیب عبارتند از

$$\mathbb{E}(T|Y=y)\quad=\delta(\alpha)(y-\xi)+\frac{\phi(\frac{\delta(\alpha)(y-\xi)}{\omega\sqrt{۱-\delta^۲(\alpha)}})}{\Phi(\frac{\delta(\alpha)(y-\xi)}{\omega\sqrt{۱-\delta^۲(\alpha)}})}\,\omega\sqrt{۱-\delta^۲(\alpha)}, \tag{۴}$$

و

$$\begin{aligned}\mathbb{E}(T^{۲}|Y=y) &= \delta^{۲}(\alpha)(y-\xi)^{۲}+\omega^{۲}(۱-\delta^{۲}(\alpha)) \\ &+ \frac{\phi(\frac{\delta(\alpha)(y-\xi)}{\omega\sqrt{۱-\delta^{۲}(\alpha)}})}{\Phi(\frac{\delta(\alpha)(y-\xi)}{\omega\sqrt{۱-\delta^{۲}(\alpha)}})}\delta(\alpha)(y-\xi)\,\omega^{۲}(۱-\delta^{۲}(\alpha)). \qquad (۵)\end{aligned}$$

در بسیاری از کاربردهای عملی، توزیع‌های مشاهدات علاوه بر چوله بودن، چندنمایی نیز هستند. این وضعیت زمانی پیش می‌آید که جامعه ناهمگن بوده و از زیرجوامعی با ویژگی‌های مختلف تشکیل شده است. در این حالت، می‌توان از مدل رگرسیونی چوله‌نرمال با متغیرهای پاسخ آمیخته به‌عنوان جایگزین مدل سنتی استفاده کرد. چنانچه متغیر پاسخ دارای توزیع آمیخته g مولفه‌ای باشد، مدل رگرسیونی را می‌توان به‌صورت

$$Y_i|\boldsymbol{x}_i \sim \sum_{j=۱}^{g}\eta_j f_{SN}(y_i|\boldsymbol{\theta}_j(\boldsymbol{x}_i)),\ \ i=۱,\ldots,n, \qquad (۶)$$

نوشت، که در آن $\boldsymbol{\theta}_j(\boldsymbol{x}_i)=(\xi_j(\boldsymbol{x}_i),\omega_j^{۲},\alpha_j)$، $\xi_j(\boldsymbol{x}_i)=h(\boldsymbol{x}_i,\boldsymbol{\beta}_j)$، $\boldsymbol{x}_i'=(x_{i۱},\cdots,x_{ip})$ بردار متغیرهای تبیینی، $\boldsymbol{\beta}_j=(\beta_{j۱},\cdots,\beta_{jp})$ بردار ضرایب رگرسیونی زیرجامعه jام و $\boldsymbol{\eta}=(\eta_۱,\ldots,\eta_p)$ بردار ضرایب آمیختگی را نشان می‌دهد. در این صورت، مجموعه پارامترهای مدل را می‌توان به‌صورت $\boldsymbol{\Psi}=(\eta_۱,\cdots,\eta_g,\boldsymbol{\theta}_۱,\cdots,\boldsymbol{\theta}_g)$ نمایش داد. در رهیافت فراوانی‌گرا، برآورد پارامترها در توزیع‌های آمیخته متناهی معمولاً از طریق الگوریتم EM صورت می‌پذیرد. برای این منظور، مشاهداتی را که از یک توزیع آمیخته متناهی به‌دست می‌آیند، از آن جهت که معلوم نیست هر مشاهده به کدام مؤلفه از توزیع آمیخته منتسب است، مشاهدات ناقص می‌نامند. سپس، مسأله را به کمک مجموعه‌ای از متغیرهای نشان‌گر در قالب داده‌های کامل بازنویسی نموده و مراحل E (امیدگیری) و M (ماکسیمم‌سازی) به صورت متوالی تکرار می‌شوند. بر این اساس، بردار نشان‌گر متناظر با مشاهده iام، $\boldsymbol{Z}_i=(Z_{i۱},\cdots,Z_{ig})$، دارای توزیع چندجمله‌ای $M(۱,\eta_۱,\cdots,\eta_g)$ است. بعلاوه، چون هر زیرجامعه دارای یک توزیع چوله‌نرمال است، روابط

$$[T_i|Z_{ij}=۱]\sim TN(۰,\omega_j^{۲}),\qquad [T_i|Y_i=y_i,Z_{ij}=۱]\sim TN(\mu_{T_{ij}},\sigma_{T_j}^{۲}), \qquad (۷)$$

برقرارند، که در آن‌ها $\mu_{T_{ij}}=\delta(\alpha_j)(y_i-\mathbf{x_i}'\boldsymbol{\beta}_j)$ و $\sigma_{T_j}=\omega_j\sqrt{۱-\delta^{۲}(\alpha_j)}$. بر این اساس،

توابع درستنمایی و لگ‌درستنمایی داده‌های کامل را می‌توان به‌ترتیب به‌صورت

$$
\begin{aligned}
L_c(\mathbf{\Psi},\mathbf{T}|\boldsymbol{y}) &= \prod_{i=1}^{n}\prod_{j=1}^{g}\{[\frac{\eta_j}{\pi\omega_j^2\sqrt{1-\delta^2(\alpha_j)}}]^{z_{ij}}\} \\
&\times \prod_{j=1}^{g}\{\exp\{\frac{-1}{2\omega_j^2(1-\delta^2(\alpha_j))}\sum_{i=1}^{n} z_{ij}[(y_i-\mathbf{x_i}'\boldsymbol{\beta}_j)^2 \\
&- 2t_i\delta(\alpha_j)(y_i-\mathbf{x_i}'\boldsymbol{\beta}_j)+t_i^2]\}\}, \qquad (8)\\
\ell_c(\mathbf{\Psi},\mathbf{T}|\boldsymbol{y}) &= \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\{\log(\eta_j)-\log(\pi\omega_j^2)-\frac{1}{2}\log(1-\delta^2(\alpha_j))\} \\
&- \sum_{g=1}^{g}\frac{1}{2\omega_j^2(1-\delta^2(\alpha_j))}[\sum_{i=1}^{n} z_{ij}(y_i-\mathbf{x_i}'\boldsymbol{\beta}_j)^2 \\
&- 2\delta(\alpha_j)\sum_{i=1}^{n} z_{ij}t_i(y_i-\mathbf{x_i}'\boldsymbol{\beta}_j)+\sum_{i=1}^{n} z_{ij}t_i^2]\},
\end{aligned}
$$

نوشت، که در آن $\boldsymbol{y} = (y_1,\ldots,y_n)'$. از این رو، برآورد ماکسیمم درستنمایی پارامترها در تکرار $(k+1)$ام الگوریتم به‌صورت

$$
\begin{aligned}
\hat{\eta}_j^{(k+1)} &= \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}, \\
\hat{\boldsymbol{\beta}}_j^{(k+1)} &= (\boldsymbol{X}'\boldsymbol{Z}_j^{*(k)}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}_j^{*(k)}\boldsymbol{y}-\delta(\hat{\alpha}_j^{(k)})\boldsymbol{X}'\hat{\mathbf{S}}_{1j}^{(k)}), \\
\hat{\omega}_j^{2(k+1)} &= \frac{(\boldsymbol{y}-\boldsymbol{X}'\hat{\boldsymbol{\beta}}_j^{(k+1)})'\boldsymbol{Z}_j^{*(k)}(\boldsymbol{y}-\boldsymbol{X}'\hat{\boldsymbol{\beta}}_j^{(k+1)})}{2(1-\delta^2(\hat{\alpha}_j^{(k)}))\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}} \\
&+ \frac{-2\delta(\hat{\alpha}_j^{(k)})(\boldsymbol{y}-\boldsymbol{X}'\hat{\boldsymbol{\beta}}_j^{(k+1)})'\hat{\mathbf{S}}_{1j}^{(k)}+\boldsymbol{J}'\mathbf{S}_{2j}^{(k)}}{2(1-\delta^2(\hat{\alpha}_j^{(k)}))\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}}, \\
\hat{\boldsymbol{\alpha}}^{(k+1)} &= \arg\max\sum_{i=1}^{n}\log(\sum_{j=1}^{g}\hat{\eta}_j^{(k+1)}f_{SN}(y_i|\hat{\boldsymbol{\beta}}_j^{(k+1)},\hat{\omega}_j^{2(k+1)},\alpha_j)),
\end{aligned}
$$

به‌دست می‌آیند، که در آن‌ها $\boldsymbol{Z}_j^{*(k)} = diag(\hat{z}_{1j}^{(k)},\hat{z}_{2j}^{(k)},\cdots,\hat{z}_{nj}^{(k)})$، $\mathbf{S}_{1j}^{(k)} = (\hat{s}_{11j}^{(k)},\hat{s}_{12j}^{(k)},\cdots,\hat{s}_{1nj}^{(k)})'$،

$\mathbf{S}_{2j}^{(k)} = (\hat{s}_{21j}^{(k)}, \hat{s}_{22j}^{(k)}, \cdots, \hat{s}_{2nj}^{(k)})'$، $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]'$، $\mathbf{J} = (1, \ldots, 1)'$ و

$$\begin{aligned}
\hat{z}_{ij}^{(k)} &= \mathbb{E}_{\hat{\boldsymbol{\Psi}}^{(k)}}(Z_{ij}|\boldsymbol{y}) = \frac{\hat{\eta}_j^{(k)} f_{SN}(y_i|\mathbf{x_i}'\hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\omega}_j^{2(k)}, \hat{\eta}_j^{(k)})}{\sum_{j=1}^{g} \hat{\eta}_j^{(k)} f_{SN}(y_i|\mathbf{x_i}'\hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\omega}_j^{2(k)}, \hat{\eta}_j^{(k)})}, \\
\hat{s}_{1ij}^{(k)} &= \mathbb{E}_{\hat{\boldsymbol{\Psi}}^{(k)}}(Z_{ij}T_i|\boldsymbol{y}) = \hat{z}_{ij}^{(k)}[\hat{\mu}_{T_{ij}}^{(k)} + \frac{\phi(\hat{\alpha}_j^{(k)}(\frac{y_i - \mathbf{x_i}'\hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\omega}_j^{(k)}}))}{\Phi(\hat{\alpha}_j^{(k)}(\frac{y_i - \mathbf{x_i}'\hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\omega}_j^{(k)}}))}\hat{\sigma}_{T_i}^{(k)}], \\
\hat{s}_{2ij}^{(k)} &= \mathbb{E}_{\hat{\boldsymbol{\Psi}}^{(k)}}(Z_{ij}T_i^2|\boldsymbol{y}) = \hat{z}_{ij}^{(k)}[\hat{\mu}_{T_{ij}}^{2(k)} + \hat{\sigma}_{T_i}^{2(k)} + \frac{\phi(\hat{\alpha}_j^{(k)}(\frac{y_i - \mathbf{x_i}'\hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\omega}_j^{(k)}}))}{\Phi(\hat{\alpha}_j^{(k)}(\frac{y_i - \mathbf{x_i}'\hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\omega}_j^{(k)}}))}\hat{\mu}_{T_{ij}}^{(k)}\hat{\sigma}_{T_i}^{(k)}].
\end{aligned}$$

ملاحظه می‌شود که محاسبه برآورد پارامترهای مدل رگرسیونی چوله‌نرمال آمیخته متناهی با پیچیدگی‌ها و دشواری‌های زیادی همراه است. برخی از این مشکلات حتی در حالت غیرآمیخته بودن توزیع متغیر پاسخ نیز وجود دارند که توسط بسیاری از محققان مورد بررسی قرار گرفته‌اند. پیوسی (۲۰۰۰) نشان داد که برآورد گشتاوری پارامتر چولگی همواره وجود ندارد. چیوگنا (۲۰۰۵)، فیگوئراس و همکاران (۲۰۰۷) و سارتوری (۲۰۰۶) نشان دادند که برآورد پارامتر چولگی حتی از طریق روش‌های عددی نیز به سادگی ممکن نیست. هم‌چنین، نشان داده شده است که اگر همه مشاهدات مثبت باشند، برآورد ماکسیمم درستنمایی پارامتر چولگی به دلیل کراندار نبودن تابع درستنمایی وجود ندارد (برای جزئیات بیشتر :نتون، ۲۰۰۴ و دی، ۲۰۱۰ را ببینید). این مشکلات در حالت آمیخته بودن توزیع بیشتر نمود پیدا می‌کند، به‌طوری که کران‌دار نبودن تابع درستنمایی نسبت به پارامترهای چولگی مربوط به مولفه‌های تشکیل‌دهنده یک توزیع آمیخته موجب می‌شود در همگرایی الگوریتم EM مشکلاتی بروز نماید. افزون بر این مشکلات، چون خواص بهینگی برآوردگرهای ماکسیمم درستنمایی تنها در حالات بزرگ نمونه‌ای بروز می‌نمایند، رهیافت بیزی به‌طور خاص برای نمونه‌های کوچک و متوسط به‌دلیل کارکرد مطلوب‌تر جذابیت‌های بیشتری دارد. در بخش بعد، نشان داده می‌شود که رهیافت بیزی مسأله بسیار سرراست بوده و مشکلات نظری و دشواری‌های عملی رهیافت فراوانی‌گرا را ندارد.

۳ تحلیل بیزی رگرسیون چوله‌نرمال آمیخته

در این بخش تحلیل رگرسیونی چوله‌نرمال آمیخته متناهی از دیدگاه بیزی مورد بررسی قرار گرفته و یک مدل بیزی برای آن توسعه داده شده است. لازمه تحلیل بیزی آن است که پارامترهای مدل متغیرهایی تصادفی تلقی شده و برای آن‌ها توزیع‌های پیشین مناسب لحاظ شود. چنانچه ریچاردسون و گرین (۱۹۹۷) اشاره نموده‌اند، در تحلیل بیزی مدل‌های آمیخته متناهی نمی‌توان تمام توزیع‌های پیشین را ناآگاهی بخش در نظر گرفت. زیرا همیشه این امکان وجود دارد که به یک یا چند زیرجامعه هیچ مشاهده‌ای تعلق نگیرد و از این‌رو مشاهدات درباره آن مولفه‌ها آگاهی بخش نباشند، که این امر موجب ناسره شدن توزیع پسین خواهد شد. در این مقاله، به پیروی از گلمن و همکاران (۲۰۰۸) برای ضرایب رگرسیونی و پارامتر واریانس مربوط به هر مؤلفه از توزیع آمیخته، به‌ترتیب توزیع‌های پیشین نرمال و وارون گاما در نظر گرفته شده‌اند. برای ضرایب آمیختگی، $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_2)$، نیز از توزیع پیشین دیریکله استفاده شده است. این انتخاب یکی از انتخاب‌های مرسوم در ادبیات توزیع‌های آمیخته محسوب می‌شود. بر این اساس، برای پارامترهای مدل توزیع‌های پیشین

$$
\begin{aligned}
\boldsymbol{\beta}_j &\sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma_j}), \qquad j = 1, \cdots, g, \\
\omega_j^{-2}|b &\sim \Gamma(a, b), \qquad j = 1, \cdots, g; \qquad a, b \in \mathcal{R}^+, \\
b &\sim \Gamma(v_1, v_2), \qquad v_1, \in \mathcal{R}^+, \\
\boldsymbol{\eta} &\sim Dirichlet(h_1, \cdots, h_g), \qquad h_1, \ldots, h_g \in \Re^+ \\
\delta(\alpha_j) &\sim U(-1, 1), \qquad j = 1, \cdots, g.
\end{aligned} \tag{9}
$$

در نظر گرفته شده است، که در آن‌ها $\boldsymbol{\mu}_j$، $\boldsymbol{\Sigma_j}$، v_1، v_2، a و $(h_1, \ldots, h_g)$، ابرپارامترهای مدل بیزی هستند و توسط تحلیل‌گر تعیین یا برآورد می‌شوند. با در نظر گرفتن یک توزیع پیشین برای ابرپارامتر b، در توزیع پیشین وارون گاما، تلاش شده است استواری بیشتری به استنباط‌های بیزی حاصل القاء شود. هم‌چنین، به‌منظور پرهیز از پیش‌داوری درباره سهم مولفه‌های توزیع آمیخته از نمونه تصادفی، معمولاً ابرپارامترهای توزیع پیشین دیریکله به‌صورت $h_1, \ldots, h_g = h$ برابر در نظر گرفته می‌شوند. این توزیع‌های پیشین انتخاب‌های بی‌طرفانه‌ای محسوب می‌شوند که برخی از مزایای مدل‌سازی بیزی برپایه توزیع‌های پیشین مزدوج را نیز به‌همراه دارند. این انتخاب‌ها تحلیل بیزی را در حالتی مد نظر قرار می‌دهند

که اطلاعات پیشین قابل‌توجه و قوی درباره پارامترهای مدل در دسترس تحلیل‌گر قرار ندارد. اگر اطلاعات پیشین قابل توجهی در دسترس باشند، می‌توان توزیع‌های پیشین آگاهی‌بخش‌تری را مد نظر قرار داد. برای ملاحظه جزئیات بیشتر درباره پیشین‌های انتخاب شده، به‌عنوان نمونه، ریچاردسون و گرین (۱۹۹۷) و برناردی و همکاران (۲۰۱۲) را ملاحظه نمائید.

برای تعیین ابرپارامترهای توزیع پیشین نرمال، داده‌ها را بر اساس میزان شباهت آن‌ها به g گروه تقسیم کرده و از برآورد ضرایب رگرسیونی و ماتریس کواریانس متناظر با آن‌ها، برای تعیین ابر پارامترهای هر گروه استفاده می‌شود. با توجه به تابع درستنمایی داده‌های کامل، رابطه (۸)، توزیع پسین داده‌های کامل پس از انجام محاسبات جبری، به‌صورت

$$\begin{aligned}
&\pi(\boldsymbol{\Psi},\mathbf{T}|\boldsymbol{y}) = L_c(\boldsymbol{\Psi},\mathbf{T}|\boldsymbol{y}) \times \pi(\boldsymbol{\Psi}) \propto \prod_{j=1}^{g}\prod_{i=1}^{n}\{\frac{\eta_j}{\pi\omega_j^2\sqrt{1-\delta^2(\alpha_j)}}\\
&\times \exp\{-\frac{1}{2\omega_j^2(1-\delta^2(\alpha_j))}[(y_i-\mathbf{x}_i'\boldsymbol{\beta}_j)^2 - 2\delta(\alpha_j)t_i(y_i-\mathbf{x}_i'\boldsymbol{\beta}_j)+t_i^2]\}\}^{Z_{ij}}\\
&\times \prod_{j=1}^{g}\{|\boldsymbol{\Sigma_j}|^{-p/2}\exp\{-\frac{1}{2}(\boldsymbol{\beta}_j-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma_j}^{-1}(\boldsymbol{\beta}_j-\boldsymbol{\mu}_j)\}b^a(\omega_j^{-2})^{a-1}e^{-b\omega_j^{-2}}\eta_j^h\}\\
&\times b^{v_1-1}e^{-v_2 b},
\end{aligned}$$

به‌دست می‌آید. با توجه به پیچیدگی توزیع پسین (۱۰) برآوردگرهای بیزی پارامترهای مدل، فاقد صورت بسته هستند. از این رو، لازم است از روش‌های مونت‌کارلو زنجیر مارکوفی برای نمونه‌گیری از توزیع پسین و تقریب کمیت‌های پسینی مورد علاقه استفاده نمود. برای این منظور، در ادامه یک الگوریتم گیبز برای نمونه‌گیری از توزیع پسین توسعه داده شده است. لازمه استفاده از الگوریتم گیبز شناخت توزیع‌های پسین شرطی کامل پارامترها است. این توزیع‌ها پس از انجام محاسبات جبری به‌صورت

$$\begin{aligned}
&\mathbf{Z_i}^{(k+1)}|others \sim M(1,\eta_1^*,\cdots,\eta_g^*),\\
&T_i^{(k+1)}|others \sim TN(\delta(\alpha_j^{(k)})(y_i-\mathbf{x}_i'\boldsymbol{\beta}_j^{(k)}),\omega_j^{2(k)}(1-\delta^2(\alpha_j^{(k)}))),\\
&b^{(k+1)}|others \sim \Gamma(v_1+ga, v_2+\sum_{j=1}^{g}\omega_j^{-2(k)}),\\
&\boldsymbol{\eta}^{(k+1)}|others \sim Dirichlet(h+n_1^{(k+1)},\cdots,h+n_g^{(k+1)}),\\
&\boldsymbol{\beta}_j^{(k+1)}|others \sim N(\boldsymbol{M}_j^{(k+1)},\boldsymbol{S}_j^{(k+1)}),
\end{aligned}$$

$$\omega_j^{-2^{(k+1)}}|others \sim \Gamma(a+n_j^{(k+1)}, b^{(k+1)}+\gamma),$$
$$\pi(\boldsymbol{\delta}|others) \propto \prod_{i=1}^{n}\prod_{j=1}^{g}\{(1-\delta^2(\alpha_j))^{-\frac{1}{2}}\exp\{-\frac{1}{2\omega_j^{2^{(k+1)}}(1-\delta^2(\alpha_j))}[$$
$$(y_i-\mathbf{x}_i'\boldsymbol{\beta}_j^{(k+1)})^2-2\delta(\alpha_j)t_i^{(k+1)}(y_i-\mathbf{x}_i'\boldsymbol{\beta}_j^{(k+1)})+t_i^{2^{(k+1)}}]\}\}^{Z_{ij}^{(k+1)}}, \quad (10)$$

به دست می‌آیند، که در آن‌ها

$$\eta_j^* = \frac{\eta_j^{(k)} f_{SN}(y_i|\mathbf{x}_i'\boldsymbol{\beta}_j^{(k)}, \omega_j^{2^{(k)}}, \alpha_j^{(k)})}{\sum_{m=1}^{g}\eta_m^{(k)} f_{SN}(y_i|\mathbf{x}_i'\boldsymbol{\beta}_m^{(k)}, \omega_m^{2^{(k)}}, \alpha_m^{(k)})},$$
$$n_j^{(k+1)} = \sum_{i=1}^{n} Z_{ij}^{(k+1)},$$
$$\mathbf{Z}_j^{*(k+1)} = diag(Z_{1j}^{(k+1)}, \cdots, Z_{nj}^{(k+1)}),$$
$$\boldsymbol{S}_j^{(k+1)} = (\frac{\mathbf{X}'\mathbf{Z}_j^{*(k+1)}\mathbf{X}}{\omega_j^{2^{(k)}}(1-\delta^2(\alpha_j^{(k)}))}+\boldsymbol{\Sigma}_j^{-1})^{-1},$$
$$\boldsymbol{M}_j^{(k+1)} = \boldsymbol{S}_j^{(k+1)}(\frac{\mathbf{X}'\mathbf{Z}_j^{*(k+1)}\mathbf{y}-\delta(\alpha_j^{(k)})\mathbf{X}'\mathbf{Z}_j^{*(k+1)}\mathbf{T}^{(k+1)}}{\omega_j^{2^{(k)}}(1-\delta^2(\alpha_j^{(k)}))}+\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j),$$
$$\gamma = \frac{1}{2(1-\delta^2(\alpha_j^{(k)}))}[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_j^{(k+1)})'\mathbf{Z}_j^{*(k+1)}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_j^{(k+1)})$$
$$-2\delta(\alpha_j^{(k)})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_j^{(k+1)})'\mathbf{Z}_j^{*(k+1)}\mathbf{T}^{(k+1)}+\mathbf{T}'^{(k+1)}\mathbf{Z}_j^{*(k+1)}\mathbf{T}^{(k+1)}].$$

به‌دلیل عدم اطلاع از ثابت چگالی‌ساز در توزیع پسین شرطی کامل پارامتر $\boldsymbol{\delta} = (\delta(\alpha_1), \cdots, \delta(\alpha_g))$ در رابطه (۱۰)، برای نمونه‌گیری از این توزیع از الگوریتم متروپولیس-هاستینگز استفاده شده است. برای این منظور، به پیروی از لین و همکاران (۲۰۰۷)، ابتدا $\delta(\alpha_j)$ را به‌صورت

$$\delta^*(\alpha_j) = \log\{\frac{1+\delta(\alpha_j)}{1-\delta(\alpha_j)}\} = 2\tanh^{-1}(\delta(\alpha_j)),$$

تبدیل کرده و سپس الگوریتم متروپولیس- هاستینگز برای نمونه‌گیری از توزیع

$$g(\boldsymbol{\delta}^*) = \pi(\boldsymbol{\delta}(\boldsymbol{\delta}^*)) \prod_{j=1}^{g} J_{\delta^*(\alpha_j)},$$

به کار گرفته می‌شود، که در آن $\boldsymbol{\delta}^* = (\delta^*(\alpha_1), \cdots, \delta^*(\alpha_g))$ و $J_{\delta^*(\alpha_j)} = \frac{2e^{\delta^*(\alpha_j)}}{(1+e^{\delta^*(\alpha_j)})^2}$ ژاکوبی تبدیل است. برای این منظور، به پیروی از گلمن و همکاران (۱۹۹۶)، یک توزیع نرمال g-متغیره با میانگین $\boldsymbol{\delta}^{*(k)}$ و ماتریس کواریانس $c^2\Sigma_{\boldsymbol{\delta}^*}^{(k)}$ به‌عنوان توزیع پیشنهادی لحاظ شده است که در آن $c \approx \frac{2/4}{\sqrt{g}}$ و مقدار $\Sigma_{\boldsymbol{\delta}^*}^{(k)}$ توسط وارون ماتریس اطلاع فیشر برآورد می‌شود. پس از استخراج نمونه‌های $\delta^*(\alpha_j)$، $j = 1, \ldots, g$، می‌توان مقادیر $\delta(\alpha_j)$ و در نهایت α_j را به‌دست آورد. این تبدیل پارامتر $\delta(\alpha_j)$ با دامنه تغییرات $(-1, 1)$ را به پارامتر $\delta^*(\alpha_j)$ با دامنه تغییرات $\mathbb{R}$ تبدیل می‌کند و این امکان را فراهم می‌سازد که بتوان از توزیع پیشنهادی گلمن و همکاران (۱۹۹۶) که یک توزیع پیشنهادی کارا است، استفاده نمود. در واقع از این طریق می‌توان به یک الگوریتم متروپولیس-هاستینگس با نرخ پذیرش مطلوب برای نمونه‌گیری از توزیع پسین شرطی کامل مورد نظر دست یافت.

ذکر این نکته ضروری است که یکی از مشکلات جدی در تحلیل بیزی توزیع‌های آمیخته، وضعیتی است که از آن تحت عنوان جابجایی‌پذیری اندیس‌ها یاد می‌شود. در چنین وضعیتی، توزیع پسین توام نسبت به جایگشت پارامترها ناوردا بوده و از این رو زنجیرهای مارکوف شبیه‌سازی‌شده از توزیع پسین توام پارامترها، شناسایی‌پذیر نیستند. بر این اساس، استنباط‌های پسینی برپایه توزیع‌های پسین حاشیه‌ای و از راه محاسبه کمیت‌های پسینی مانند میانگین پسین، گمراه‌کننده خواهد بود. روش‌های مختلفی برای برخورد با مشکل جابجایی پذیری اندیس‌ها، پیشنهاد شده است که خواننده علاقه‌مند می‌تواند برای ملاحظه برخی از این روش‌ها، استفنس (۲۰۰۰) و جاسارا و همکاران (۲۰۰۵) را ملاحظه نماید. در این مقاله، برای مواجهه با این مشکل از بسته کتابخانه‌ای $label.switching$ در نرم‌افزار R که توسط پاپاستامولیس (۲۰۱۵) توسعه داده شده است، بهره برده‌ایم. این بسته نرم‌افزاری مجموعه‌ای از جایگشت‌ها را برای مرتب‌سازی مجدد زنجیرهای مارکوف شبیه‌سازی‌شده از توزیع پسین، فراهم می‌سازد. هرگونه استنباط پسینی بر پایه این زنجیرهای مجدداً مرتب شده، صورت می‌پذیرد.

۴ مطالعه شبیه‌سازی

در این بخش برای ارزیابی رهیافت بیزی پیشنهادی یک مدل رگرسیونی آمیخته دومؤلفه‌ای به‌صورت

$$\begin{aligned} Y_i|\mathbf{x}_i &\sim \eta f_{SN}(\beta_{۱۱}x_{i۱}+\beta_{۱۲}x_{i۲},\omega_۱^۲,\alpha_۱) \\ &+ (۱-\eta)f_{SN}(\beta_{۲۱}x_{i۱}+\beta_{۲۲}x_{i۲},\omega_۲^۲,\alpha_۲), \end{aligned} \tag{۱۱}$$

$i = ۱,۲,\cdots,n$، در نظر گرفته شده است. با فرض $(\beta_{۱۱},\beta_{۱۲}) = (۹,۱۵)'$، $(\beta_{۲۱},\beta_{۲۲}) = (-۱,-۶)'$، $\omega_۱^۲ = ۰/۵$، $\omega_۲^۲ = ۰/۵$ و $\eta = ۰/۵$، مقادیر متغیرهای پاسخ از توزیع (۱۱) شبیه‌سازی شده‌اند. به‌منظور لحاظ نمودن وضعیت‌های مختلف ممکن در ارزیابی رهیافت بیزی پیشنهادی، برای مؤلفه اول-دوم توزیع آمیخته (۱۱)، چهار وضعیت مختلف چوله به راست-چوله به راست، چوله به راست-متقارن، چوله به راست-چوله به چپ، و متقارن-متقارن، به‌ترتیب برای مولفه‌های اول و دوم توزیع آمیخته، در نظر گرفته شده است.

برای ارزیابی تأثیر اندازه نمونه بر کارایی برآوردگرهای بیزی، مطالعه شبیه‌سازی به‌ازای مقادیر مختلف اندازه نمونه ۱۰، ۲۰، ۴۰ و ۸۰ تکرار شده است. چون رهیافت پیشنهادی بیزی است، از در نظر گرفتن اندازه نمونه‌های بزرگتر پرهیز شده است، چرا که مزیت اصلی رهیافت بیزی عمدتاً مربوط به اندازه‌های کوچک نمونه‌ای است و برای نمونه‌های بزرگ به‌دلیل غلبه درستنمایی بر پیشین، نتایج دو رهیافت بیزی و فراوانی‌گرا مشابه خواهند بود. به‌منظور لحاظ نمودن عدم قطعیت و تغییرپذیری حاکم بر فرایند تولید نمونه‌های تصادفی، تمام محاسبات ۱۰۰۰ بار تکرار و مقادیر جذر میانگین توان‌های دوم خطای برآوردگرهای ضرایب رگرسیونی، برای وضعیت‌های چهارگانه یادشده محاسبه و در جداول ۱ الی ۴ ارائه شده‌اند.

برای اطمینان از همگرایی زنجیرهای مارکوف تولید شده به توزیع مانای متناظر خود، از معیار پیشنهادی گلمن و روبین (۱۹۹۲) استفاده شده است. برای این منظور تابع $gelman.diag$ در بسته کتابخانه‌ای $Coda$ (پلومر و همکاران، ۲۰۰۶) از نرم‌افزار R (گروه مرکزی R، ۲۰۱۳) مورد استفاده قرار گرفته است. لازم به ذکر است که زمان اجرای الگوریتم‌ها بسته به تعداد مولفه‌های توزیع آمیخته، مقادیر مختلف ضرایب آمیختگی و اندازه نمونه‌های مختلف، متفاوت است. به عنوان مثال، در یک توزیع آمیخته دومولفه‌ای، به‌ازای ضریب آمیختگی ۰/۵ و اندازه نمونه ۴۰، زمان همگرایی زنجیر مارکوف تولید شده از توزیع پسین در یک بار اجرای الگوریتم گیبز، اندکی کمتر از ۲۰ دقیقه بوده است.

با توجه به مقادیر جذر میانگین توان‌های دوم خطای مربوط به ضرایب رگرسیونی، ملاحظه می‌شود

جدول ۱. مقادیر جذر میانگین توان دوم خطای برآوردگرهای ضرایب رگرسیونی برای مدل‌های رقیب در وضعیت متقارن-متقارن ($\alpha_۱ = ۰, \alpha_۲ = ۰$).

	چوله‌نرمال آمیخته			نرمال آمیخته		
n	۲۰	۴۰	۸۰	۲۰	۴۰	۸۰
$\beta_{۱۱}$	۰/۰۰۰۱	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$
$\beta_{۱۲}$	۰/۰۰۰۱	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$
$\beta_{۲۱}$	۰/۰۰۰۳	$< ۱۰^{-۱۲}$	۰/۰۰۰۱	۰/۰۰۰۷	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$
$\beta_{۲۲}$	۰/۰۰۰۴	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	۰/۰۰۰۵	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$
$\omega_۱^۲$	۰/۰۰۲۳	۰/۰۰۸۹	۰/۰۰۳۶	۰/۰۰۹۲	۰/۰۰۰۳	۰/۰۰۰۸
$\omega_۲^۲$	۰/۰۰۲۰	۰/۰۰۹۳	۰/۰۰۴۱	۰/۰۰۱۳	۰/۰۰۲۹	۰/۰۰۰۶
$\alpha_۱$	۰/۰۰۰۱	$< ۱۰^{-۱۲}$	۰/۰۰۰۴	-	-	-
$\alpha_۲$	۰/۰۰۰۴	۰/۰۰۰۱	۰/۰۰۰۱	-	-	-
η	۰/۰۰۱۵	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	۰/۰۰۲۰	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$

جدول ۲. مقادیر جذر میانگین توان دوم خطای برآوردگرهای ضرایب رگرسیونی برای مدل‌های رقیب در وضعیت چوله به راست-متقارن ($\alpha_۱ = ۴, \alpha_۲ = ۰$).

	چوله‌نرمال آمیخته			نرمال آمیخته		
n	۲۰	۴۰	۸۰	۲۰	۴۰	۸۰
$\beta_{۱۱}$	۰/۰۱۵۳	۰/۰۰۶۴	۰/۰۰۱۹	۰/۰۷۵۴	۰/۰۵۳۲	۰/۰۸۱۰
$\beta_{۱۲}$	۰/۰۲۵۷	۰/۰۰۸۳	۰/۰۰۲۲	۰/۱۰۸۷	۰/۰۸۵۳	۰/۰۹۰۴
$\beta_{۲۱}$	۰/۰۰۰۱	۰/۰۰۰۱	$< ۱۰^{-۱۲}$	۰/۰۰۰۴	۰/۰۰۰۱	$< ۱۰^{-۱۲}$
$\beta_{۲۲}$	۰/۰۰۰۲	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	۰/۰۰۰۲	۰/۰۰۰۱	$< ۱۰^{-۱۲}$
$\omega_۱^۲$	۰/۰۱۲۱	۰/۰۱۰۳	۰/۰۰۳۷	۰/۰۷۵۲	۰/۰۶۶۴	۰/۰۶۱۴
$\omega_۲^۲$	۰/۰۰۲۲	۰/۰۰۶۰	۰/۰۰۳۶	۰/۰۱۴۹	۰/۰۰۲۳	۰/۰۰۰۸
$\alpha_۱$	۹/۰۹۷۲	۵/۳۶۶۷	۱/۴۸۳۷	-	-	-
$\alpha_۲$	۰/۰۰۰۱	۰/۰۰۰۹	$< ۱۰^{-۱۲}$	-	-	-
η	۰/۰۰۱۴	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	۰/۰۰۲۰	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$

که وقتی جامعه ناهمگن بوده و از زیرجوامعی با ویژگی‌های متفاوت تشکیل شده است، توزیع‌های آمیخته متناهی به خوبی ناهمگنی و تمایز بین زیرجوامع را مدنظر قرار می‌دهند. بعلاوه، هنگامی که توزیع‌های مربوط به هر زیرجامعه دارای ساختار نامتقارن هستند، امکان مدل‌بندی این عدم تقارن به کمک توزیع‌های آمیخته چوله‌نرمال به خوبی فراهم آورده می‌شود. کاهش مقادیر جذر میانگین توان‌های دوم خطا ضرایب رگرسیونی با افزایش اندازه نمونه حاکی از سازگاری برآوردگرهای پیشنهادی است.

۵ مثال کاربردی

در این بخش نحوه کاربست رهیافت بیزی پیشنهادی را در قالب یک مثال کاربردی شرح می‌دهیم. برای این منظور، بخشی از داده‌های هزینه و درامد خانوارها، مربوط به سرشماری نفوس و مسکن سال ۱۳۸۸، مورد استفاده قرار گرفته‌است. جدول ۵ مقادیر درآمد ۳۶ خانوار ایرانی (به هزار ریال) به همراه هزینه و بعد خانوارها را نشان می‌دهد. هدف بررسی تأثیر متغیرهای تبیینی بعد و درآمد، بر متغیر پاسخ هزینه‌های خالص خانوارها است. بافت‌نگار هزینه خالص خانوارها که در شکل ۱ رسم شده است، به‌خوبی ساختار ناهمگن، دونمایی و نامتقارن مشاهدات تحت بررسی را نشان می‌دهد.

با این وجود نمی‌توان در مورد تعداد مولفه‌های توزیع آمیخته تنها بر پایه شکل توزیع قضاوت نمود، چرا که در بسیاری حالات نماهای مختلف یک توزیع آمیخته متناهی دارای هم‌پوشانی بوده و به‌راحتی قابل تمایز نیستند. در اغلب کاربردها تعداد مولفه‌های یک توزیع آمیخته متناهی مشخص نیست و می‌بایست به‌عنوان یک پارامتر بر اساس مشاهدات موجود برآورد شود. در طی دهه‌های گذشته روش‌های مختلفی برای این منظور پیشنهاد شده است. یکی از عمومی‌ترین روش‌ها در این زمینه، روش مبتنی بر معیارهای انتخاب مدل است. در این روش‌ها مدل‌هایی با تعداد مولفه‌های مختلف به مشاهدات برازش و سپس تعداد مولفه‌ها به‌صورت $\hat{g} = arg\min_{g\in G} C(\hat{\mathbf{\Psi}}_g, g)$ برآورد می‌شود، که در آن G مجموعه مقادیر ممکن برای تعداد مولفه‌ها، $\hat{\mathbf{\Psi}}$ برآوردگر پارامترهای مدل با فرض آنکه مدل g مولفه دارد و $C(\cdot)$ یک معیار انتخاب مدل مانند معیار اطلاع آکائیک (AIC)، شوارتز (BIC) و هانان-کوئین (HQC) یا

جدول ۳. مقادیر جذر میانگین توان دوم خطای برآوردگرهای ضرایب رگرسیونی برای مدل‌های رقیب در وضعیت چوله به راست-چوله به چپ ($\alpha_۱ = ۴, \alpha_۲ = -۴$).

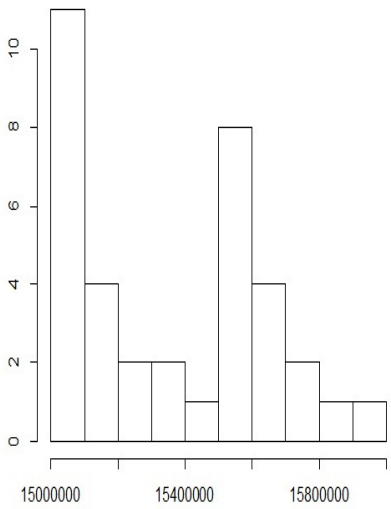
	چوله‌نرمال آمیخته			نرمال آمیخته		
n	۲۰	۴۰	۸۰	۲۰	۴۰	۸۰
$\beta_{۱۱}$	۰/۰۲۵۹	۰/۰۰۵۸	۰/۰۰۱۰	۰/۰۸۸۷	۰/۰۵۶۰	۰/۰۷۵۶
$\beta_{۱۲}$	۰/۰۴۱۲	۰/۰۰۵۶	۰/۰۰۱۴	۰/۱۲۶۳	۰/۰۵۳۴	۰/۰۶۴۵
$\beta_{۲۱}$	۰/۰۸۱۶	۰/۰۰۵۶	۰/۰۰۱۵	۰/۳۸۴۶	۰/۰۵۵۵	۷۶۴ ۰/۰
$\beta_{۲۲}$	۰/۰۰۷۷	۰/۰۰۵۲	۰/۰۰۱۴	۰/۰۱۹۲	۰/۰۵۲۶	۰/۰۶۵۹
$\omega_۱^۲$	۰/۰۲۶۳	۰/۰۰۱۴	۰/۰۰۳۸	۰/۰۸۶۶	۰/۰۶۳۵	۰/۰۵۸۹
$\omega_۲^۲$	۰/۰۲۱۱	۰/۰۰۳۳	۰/۰۰۴۸	۰/۰۷۹۴	۰/۰۶۲۶	۰/۰۶۱۳
$\alpha_۱$	۱۰/۰۱۳۵	۵/۳۰۷۹	۱/۲۶۱۶	-	-	-
$\alpha_۲$	۹/۹۷۲۷	۵/۰۲۵۶	۱/۱۷۶۹	-	-	-
η	۰/۰۰۱۳	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	۰/۰۰۱۸	$< ۱۰^{-۱۲}$	۰/۰۰۲۹

جدول ۴. مقادیر جذر میانگین توان دوم خطای برآوردگرهای ضرایب رگرسیونی برای مدل‌های رقیب در وضعیت چوله به راست-چوله به راست ($\alpha_۱ = ۴, \alpha_۲ = ۴$).

	چوله‌نرمال آمیخته			نرمال آمیخته		
n	۲۰	۴۰	۸۰	۲۰	۴۰	۸۰
$\beta_{۱۱}$	۰/۰۶۰۸	۰/۰۱۱۷	۰/۰۰۲۵	۰/۱۴۶۷	۰/۰۸۷۳	۰/۰۷۲۴
$\beta_{۱۲}$	۰/۰۲۲۰	۰/۰۱۱۰	۰/۰۰۱۹	۰/۰۵۲۸	۰/۰۸۱۰	۰/۰۵۹۲
$\beta_{۲۱}$	۰/۰۲۲۵	۰/۰۰۹۹	۰/۰۰۲۱	۰/۰۴۸۰	۰/۰۸۱۶	۰/۰۷۶۱
$\beta_{۲۲}$	۰/۰۴۵۸	۰/۰۱۰۰	۰/۰۰۲۲	۰/۱۱۲۳	۰/۰۷۹۶	۰/۰۵۹۹
$\omega_۱^۲$	۰/۰۳۳۵	۰/۰۱۶۵	۰/۰۰۶۱	۰/۰۹۶۰	۰/۰۶۸۲	۰/۰۶۲۶
$\omega_۲^۲$	۰/۰۳۰۳	۰/۰۱۶۹	۰/۰۰۷۷	۰/۰۹۴۳	۰/۰۷۳۳	۰/۰۶۶۸
$\alpha_۱$	۱۱/۴۶۲۲	۶/۰۸۶۲	۱/۷۹۵۹	-	-	-
$\alpha_۲$	۱۱/۵۷۴۰	۵/۹۸۹۳	۱/۸۷۴۲	-	-	-
η	۰/۰۰۱۵	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$	۰/۰۰۲۱	$< ۱۰^{-۱۲}$	$< ۱۰^{-۱۲}$

جدول ۵. مقادیر درآمد ۳۶ خانوار ایرانی (به ۱۰۰۰ ریال) به همراه هزینه و بعد خانوارها.

ردیف	هزینه	بعد خانوار	درآمد	ردیف	هزینه	بعد خانوار	درآمد
۱	۱۵۰۱۴	۲	۱۹۰۶۰	۱۹	۱۵۰۰۴۶/۴	۳	۲۰۰۷۰
۲	۱۵۰۶۲/۴	۳	۹۸۵۰	۲۰	۱۵۶۱۰/۴۸	۴	۲۴۱۲۰
۳	۱۵۲۹۷/۷۲	۴	۳۵۰۴۰	۲۱	۱۵۱۲۳	۲	۱۵۵۸۰
۴	۱۵۵۴۳/۵۲	۲	۸۹۵۲	۲۲	۱۵۰۵۳/۹۷	۲	۱۵۳۲۹/۹۷
۵	۱۵۰۷۲	۱	۵۰۵۰۰	۲۳	۱۵۱۰۸	۱	۱۳۸۵۰۰
۶	۱۵۰۲۰	۲	۳۳۶۰۰	۲۴	۱۵۴۴۶/۸	۳	۱۷۴۳۰
۷	۱۵۵۰۸/۸	۱	۲۷۶۰۰	۲۵	۱۵۰۳۱/۲	۶	۳۵۹۵/۲
۸	۱۵۵۴۹/۲	۱	۳۰۳۰۰	۲۶	۱۵۰۳۱/۶	۲	۱۶۸۰۰
۹	۱۵۰۳۸	۱	۹۴۳۰	۲۷	۱۵۰۴۶/۶	۷	۴۳۹۶۰
۱۰	۱۵۰۵۴	۴	۶۱۷۶۰	۲۸	۱۵۵۸۲/۱۲	۴	۴۳۵۰۰
۱۱	۱۵۱۴۲/۲	۲	۱۳۶۵۰۰	۲۹	۱۵۶۷۸	۲	۱۵۰۴۶
۱۲	۱۵۱۵۵/۳۸	۲	۱۹۶۰۰	۳۰	۱۵۳۳۵/۶۶	۱	۸۷۱۰
۱۳	۱۵۳۹۷/۷	۴	۳۴۲۲۰	۳۱	۱۵۵۶۳/۶۸	۳	۱۹۱۱۰
۱۴	۱۵۸۱۰/۲	۱	۱۶۸۰۰	۳۲	۱۵۵۵۴	۴	۲۴۷۲۰
۱۵	۱۵۵۶۰	۲	۱۶۰۰۰	۳۳	۱۵۷۰۹	۲	۲۶۰۰۰
۱۶	۱۵۵۸۲/۸۶	۲	۱۲۹۸۵/۲۶	۳۵	۱۵۶۴۶/۲۸	۱	۱۷۸۶۶/۸
۱۷	۱۵۷۰۶/۷۲	۱	۱۹۲۰۰	۳۵	۱۵۶۵۲	۲	۱۸۸۰۰
۱۸	۱۵۹۷۸	۲	۳۴۲۱۲	۳۶	۱۵۲۴۰/۵۴	۳	۴۳۵۹۰



شکل ۱. بافت‌نگار هزینه خالص خانوارهای شهری.

DIC بیزی است. سه معیار اول به‌ترتیب به‌صورت

$$\begin{aligned} AIC &= -۲\ell(\hat{\boldsymbol{\theta}}) + ۲k, \\ BIC &= -۲\ell(\hat{\boldsymbol{\theta}}) + k\log n, \\ HQC &= -۲\ell(\hat{\boldsymbol{\theta}}) + ۲k\log(\log n), \end{aligned}$$

محاسبه می‌شوند، که در آن‌ها k تعداد پارامترهای هر مدل و n اندازه نمونه را نشان می‌دهد. معیار DIC بیزی به‌صورت

$$DIC = \bar{D} - D(\bar{\boldsymbol{\theta}}), \qquad (۱۲)$$

تعریف می‌شود، که در آن $D(\boldsymbol{\theta}) = -۲\ell(\boldsymbol{\theta})$ را انحراف گویند، $\bar{D} = E(D(\boldsymbol{\theta})|\boldsymbol{y})$ و $\bar{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{y})$. برای ملاحظه جزئیات بیشتر درباره نحوه محاسبه و تعبیر معیار DIC، اشپیگلهارتر و همکاران (۲۰۰۲)

را ببینید. جدول ۶ مقادیر معیارهای اطلاع مختلف را برای مدل چوله‌نرمال آمیخته به‌ازای تعداد مولفه‌های مختلف، نشان می‌دهد. همان‌طور که ملاحظه می‌شود، مدل چوله‌نرمال آمیخته دومولفه‌ای، بر اساس همه معیارهای اطلاع یادشده، در مقایسه با سایر مدل‌ها از حمایت بیشتری از سوی مشاهدات برخوردار است. از این‌رو، برای مدل‌بندی متغیر پاسخ از یک توزیع چوله‌نرمال آمیخته دومؤلفه‌ای به‌صورت

$$\begin{aligned} Y_i|\mathbf{x}_i &\sim \eta f_{SN}(\beta_{11}x_{i1}+\beta_{12}x_{i2},\omega_1^2,\alpha_1) \\ &+ (1-\eta) f_{SN}(\beta_{21}x_{i1}+\beta_{22}x_{i2},\omega_2^2,\alpha_2), \end{aligned} \tag{13}$$

$i = 1, 2, \cdots, n$، استفاده شده است. پارامترهای این مدل بر اساس رهیافت بیزی پیشنهادی برآورد و در جدول ۷ ارائه شده‌اند. مدل فراوانی‌گرایانه نرمال آمیخته نیز به‌عنوان مدل سنتی رقیب به داده‌ها برازش داده شده است، تا اثر عدم تقارن مشاهدات بر کارایی مدل‌های رقیب به‌صورت دقیق‌تر مورد مطالعه قرار بگیرد. به منظور ارزیابی مدل بیزی پیشنهادی و مقایسه آن با مدل رقیب فراوانی‌گرا، مقادیر معیارهای اطلاع AIC، BIC و HQC که ذاتاً معیارهایی فراوانی‌گرا هستند، نیز برای مدل‌های مورد بحث محاسبه و در جدول ۸ ارائه شده‌اند. ملاحظه می‌شود که مدل رگرسیونی چوله‌نرمال آمیخته بیزی دومولفه‌ای حتی بر اساس معیارهای فراوانی‌گرایانه نیز، در مقایسه با مدل فراوانی‌گرا کارایی بیشتری دارد. بعلاوه، در رهیافت بیزی پیشنهادی، مشکلات معمول مدل رقیب فراوانی‌گرا مانند کران‌دار نبودن تابع درستنمایی، عدم همگرایی الگوریتم EM و بهینه نبودن نتایج برای نمونه‌های کوچک، بروز پیدا نمی‌کند. برنامه‌های نوشته‌شده برای برازش مدل رگرسیونی بیزی چوله‌نرمال آمیخته متناهی، در نشانی اینترنتی $https://sites.google.com/site/datasetandprograms$ در دسترس هستند.

بحث و نتیجه‌گیری

تحلیل رگرسیونی معمولاً با فرض متقارن بودن متغیر پاسخ صورت می‌گیرد. این در حالی است که در بسیاری از کاربردها، مشاهدات انحراف زیادی از توزیع نرمال داشته و دارای ساختار چندنمایی و چوله هستند. از این‌رو، روش‌های سنتی تحلیل رگرسیونی به نتایج گمراه‌کننده‌ای منجر می‌شوند. در چنین شرایطی، توزیع‌های آمیخته متناهی می‌توانند به خوبی ناهمگنی و تمایز بین زیرجوامع را مدل‌بندی نمایند. علی‌رغم ویژگی‌های جالب توزیع چوله‌نرمال، برآورد پارامتر چولگی در رهیافت فراوانی‌گرا با مشکلات خاصی همراه است. با توجه به دشواری‌های ذاتی برآورد پارامترها در توزیع‌های آمیخته و این که خواص بهینه برآوردگرهای ماکسیمم

جدول ۶. مقادیر معیارهای انتخاب مدل مختلف برای مدل چوله‌نرمال آمیخته بیزی با تعداد مولفه‌های متفاوت.

تعداد مولفه‌ها	تعداد پارامتر	BIC	AIC	HQC	DIC
۲	۹	۲۵۱/۰۰	۲۳۶/۷۴	۲۸۳/۲۴	۲۱۴/۶۷
۳	۱۴	۲۷۴/۴۴	۲۵۲/۲۷	۳۲۴/۶۱	۲۱۵/۷۵
۴	۱۹	۲۹۶/۰۹	۲۶۴/۰۱	۳۶۴/۱۸	۲۳۰/۸۵

جدول ۷. مقادیر برآورد پارامترهای دو مدل چوله‌نرمال آمیخته بیزی و نرمال آمیخته.

مدل آمیخته	β_{11}	β_{12}	β_{21}	β_{22}	ω_1^2	ω_2^2	α_1	α_2
چوله‌نرمال	۱/۶۲۴۰	۰/۱۹۰۴	۲/۱۲۳۹	۰/۶۰۸۹	۵۸/۴۲	۰/۳۴	۲/۲۸	۰/۰۹
نرمال	۲/۰۰۸۰	۰/۲۵۶۴	۲/۴۷۴۳	۰/۶۰۱۲	۳۷/۴۰	۸/۵۳	-	-

جدول ۸. مقادیر معیارهای انتخاب مدل مختلف برای مدل چوله‌نرمال آمیخته به همراه مقادیر متناظر برای مدل نرمال آمیخته.

مدل آمیخته	تعداد مولفه‌ها	تعداد پارامترها	BIC	AIC	HQC
چوله نرمال	۲	۹	۲۵۱/۰۰	۲۳۶/۷۴	۲۸۳/۲۴
نرمال	۲	۷	۲۵۸/۸۲	۲۴۷/۷۴	۲۸۳/۹۱
	۳	۱۱	۲۶۸/۵۲	۲۵۱/۱۰	۳۰۷/۹۳
	۴	۱۵	۲۶۵/۸۳	۲۴۲/۰۷	۳۱۹/۵۸

درستنمایی تنها در حالت مجانبی تحقق می‌یابند، رهیافت فراوانی‌گرایانه تحلیل رگرسیونی تحت توزیع‌های آمیخته از کارایی لازم برخوردار نیست. نتایج حاصل از مطالعه شبیه‌سازی و مثال کاربردی نشان می‌دهد که میزان کارایی رهیافت بیزی پیشنهادی، دست‌کم برای نمونه‌های کوچک و متوسط، به‌صورت قابل توجهی از مدل فراوانی‌گرا بیشتر است. افزون بر این، رهیافت بیزی پیشنهادی از نظر محاسباتی بسیار سرراست بوده و مشکلات مختلف رهیافت فراوانی‌گرا مانند کران‌دار نبودن تابع درستنمایی، عدم همگرایی الگوریتم عدم همگرایی الگوریتم EM و بهینه نبودن نتایج برای نمونه‌های کوچک، در این رهیافت وجود ندارند.

تقدیر و تشکر

نویسندگان از داوران و سردبیر گرامی مجله برای ارائه پیشنهادهای سازنده، در راستای بهبود این پژوهش، سپاسگزاری می‌کنند.

مراجع

Aitkin, M. and Wilson, G. T. (1980), Mixture Modelse, Outliers and EM Algorithm. *Technormetrics*, **22**, 325-331.

Arellano-Valle, R. B., Castro, L. M., Genton, M. G., and Gomez, H. M. (2008), Bayesian Inference for Shape Mixtures of Skewed Distributions with Application to Regression Analysis. *Bayesian Analysis*, **3**, 513-540.

Azzalini, A. (1985), A Class of Distribution which Includes the Normal Ones, *Scandinanavian Journal of Statistics*, **12**, 171-178.

Barr, R., Donald, E. and Sherril. (1999), Mean and Variance of Truncated Normal Distribution, The American Statistician. **53**, 357-361.

Bernardi, M., Maruotti, A. and Lea, P. (2012), Skew Mixture Models for Loss Distributions: A Bayesian Approach, The American Statistician. **53**, 357-361.

Cancho, V. G., Dey, K. D., Lachos, V. H., and Andrade, M. (2010a), Bayesian Nonlinear Regression Models with Scale Mixtures of Skew Normal Distributions: Estimation and Case influence diagnostics *Computational Statistics and Data Analysis*, **55**, 588 -602.

Cancho, V. G., Lachos, V. H., and Ortega, E. M. M. (2010b), A Nonlinear Regression Model with Skew-Normal Errors, *Statistical Papers*, **51**, 547 -558.

Chiogna, M. (2005), A Note on the Asymptotic Distribution of the Maximum Likelihood Estimator for the Scalar for the Skew-Normal Disrtibution, *Statistical Methods and Applications*, **14**, 331 -347.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), Maximum Likelihood with Incomplete data via the E-M algorithm. *Journal of the Royal Statistical Society*, **39**, 1-38.

DeSarbo, W. S. and Corn, W. L. (1988), A Maximum Likelihood Methodology for Clusterwise linear regression. *Journal of Classification*, **5**, 249-282.

Dey, D. (2010), Estimation of the Parameters of Skew-Normal Distribution by Apprximating the Rato of the Normal Density and Distribution Functions, PhD Thesis in *Applied Statistic*, University of California.

Figuerase, G., Puig, P. and Pewsey, A. (2007), Goodness-of-Fit Tests for the Skew Normal Distribution when the Parameters are Estimated from the Data, *Communications in Statistics: Theory and Methods*, **36**, 1735-1755.

Gelman, A., Robert, G. and Gilks, W. (1996), Efficient Metropolis Jumping Rules, In Bayesian Statistics 5 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), *Oxford University Press*, New York.

Gelman, A., Jakulin, M. Grazia Pittau, A. and Su, A. (2008), A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics*, **4**, 1360–1383.

Gelman A. and Rubin D. B. (1992), Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, **7**, 457-511.

Genton, M. G. (2004), Skew Elliptical Distribution and Thier Applications: A Journey Beyond Normality, USA, Chapman Hall/CRC.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2000), Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, **20**, 1, 50-67.

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.

Lachos, V., Ghosh, P. and Arellano–Valle, R. (2010). Likelihood Based Inference for Skew Normal/Independent Linear Mixed Model. *Statistica Sinica*, **20**, 303-322.

Lin, T. I., Lee, J. C. and Yen, S. Y. (2007), Finite Mimture Modeling Using the Skew Normal Distribution. *Statistica Sinica*, **17**,909-927.

Liu, M., Hancock, G. R., and Harring, J. R. (2011), Using Finite Mixture Modeling to Deal with Systematic Measurement Error: A Case Study. *Journal of Modern Applied Statistical*, **10**, 249-261.

Bernardi, M., Mariuotti, A. and Lea, P. (2012), Skew Mixture Models for Loss Distributions: A Bayesian Approach, *Insurance: Mathematics and Economics*, **51**, 3, 617-623.

Papastamoulis, P. (2015), Label.Switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs, *Journal of Statistical Software*, **vv**, Code Snippet II, http://www.jstatsoft.org.

Pewsey, A. (2000), Problems of Inference for Azzalini Skew - Normal Distribution. *Journal of Applied Statistics.*, **27**, 859-870.

Quandt, R. E. (1972), A New Approach to Estimating Switching Regressions. *Journal of the American Statistical Association*, **67**, 306-310.

Quandt, R. E., and Ramsey, J. B. (1978), Estimating Mixtures of Normal Distributions and Switching Regressions, *Journal of the American Statistical Association*, **73**, 730-738.

R Core Team (2013), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

Plummer, M., Best, N. Cowles, K. and Vines, K. (2006), CODA: Convergence Diagnosis and Output Analysis for MCMC, *R News*, **6**, 1, 7-11.

Richardson, S. and P. J. Green (1997), On Bayesian Analysis of Mixtures with an Unknown Number of Components (With Discusson). *Journal of the Royal Statistical Society, Series B*, **59**, 4, 731–792.

Sartori, M. (2006), Bias Prevention of Maximum Likelihood Estimation: Skew-Normal and t Distribution. *Journal of Statististical Planning and Inference*, **136**, 4259-4275.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002), Bayesian Measures of Model Complexity and Fit, *Journal of Royal Statistical Society, Series B*, **59**, 731–792.

Stephens, M. (2000), Dealing with Label Switching in Mixture Models, *Journal of Royal Statistical Society, Series B*, **82**, 4, 795-809.

Turner, T. R. (2000), Estimating the Propagation Rate of a Viral Infection of Potato Plants via Mixtures of Regressions, *Journal of Applied Statistics*, **49**, 371-384.