مجلهی بررسیهای آمار رسمی ایران
سال ۲۹، شمارهی ۲، پاییز و زمستان ۱۳۹۷، صص ۱۸۵ – ۲۰۵

استفاده از رگرسیون آمیخته چندکی برای تحلیل هزینه و درآمد خانوارهای ایرانی

عنایت بارانی و موسی گل‌علیزاده*

دانشگاه تربیت مدرس

چکیده: استقلال مولفه‌های خطا یکی از فرض‌های اساسی در مدل‌های رگرسیونی خطی مبتنی بر تغییرات میانگین متغیر پاسخ است که در صورت برقرار نبودن آن می‌توان از مدل‌های رگرسیون آمیخته خطی استفاده کرد. اما به کارگیری این مدل‌ها نیز مستلزم عدم وجود داده‌های دورافتاده است که در صورت نقص آن‌ها، مدل‌های رگرسیون آمیخته چندکی خطی با فرض توزیع لاپلاس نامتقارن برای خطاها جایگزین مناسبی خواهند بود. در این مقاله، ابتدا از طریق آزمایش شبیه‌سازی عملکرد روش‌های موجود برآوردیابی فراوانی‌گرا در اینگونه مدل‌ها مورد ارزیابی قرار گرفت. مشاهد شد که روش ترکیب شده از تقریب مربع‌بندی گاوس و الگوریتم بهینه‌سازی ناهموار برآوردهای دقیق‌تری نسبت به روش تقریبی الگوریتم تقریب تصادفی ماکسیمم‌سازی امید ریاضی ارایه می‌دهد. سپس، با توجه به همبسته بودن هزینه خانوارها و همچنین مشاهدات غیرمعمول درون هر یک از استان‌ها، مدل مورد اشاره برای تحلیل داده‌های هزینه و درآمد خانوارهای ایرانی در سال ۱۳۹۰ به کار گرفته شد. بنا به نتایج حاصل از برازش مدل انتظار می‌رود خانوارهایی که درآمد بیشتری دارند، دارای هزینه ناخالص بالاتری باشند. از طرفی سطح زیربنای محل سکونت، بُعد خانوار، جمعیت استان، خانوارهای دارای سرپرست شاغل و نحوه تصرف محل سکونت خانوارها به صورت اجاره‌ای تأثیر کمتری در هزینه ناخالص خانوارها دارند. به‌علاوه، با افزایش سن سرپرست خانوار و وجود تعداد افراد شاغل بیشتر در خانوارها، هزینه ناخالص خانوار کاهش می‌یابد.

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۸/۳/۱۲، پذیرش: ۱۳۹۸/۹/۱۷.

واژگان کلیدی: توزیع لاپلاس نامتقارن، مدل‌های رگرسیون آمیخته چندکی خطی، روش تقریب مربع‌بندی گاوس، الگوریتم ماکسیمم‌سازی امید ریاضی، داده‌های هزینه و درآمد.

۱- مقدمه

رویکرد مدل‌های آمیخته خطی، روشی برای تجزیه و تحلیل داده‌های برآمده از جامعه آماری به صورت تودرتو (خوشه‌ای، آشیانه‌ای، سلسله مراتبی و چندسطحی) است. بنا به گلداشتاین [۷] ساختار جامعه بسیاری از انواع مشاهدات مرتبط با موضوع انسان به صورت خوشه‌ای یا سلسله مراتبی است. اسنایدر و بوسکر [۲۰] نیز یادآوری کردند که ساختار جامعه آماری چندسطحی در حوزه علوم اجتماعی، جامعه‌شناسی، روان شناسی، اقتصاد و جرم‌شناسی نیز یافت می‌شود. به عنوان مثال، می‌دانیم که بنا به ویژگی‌های وراثتی، شباهت زیادی بین روحیه، شخصیت و احساسات فرزندانی که از یک خانواده هستند وجود دارد.

یکی از فرض‌های اساسی مدل رگرسیون معمولی عدم همبستگی بین مولفه‌های خطا است که در صورت برقرار نبودن آن، باید راه حلی برای مواجه با آن پیدا کرد و یا اینکه از مدل‌های آماری دیگری استفاده کرد. چنان‌چه داده‌ها دارای همبستگی درونی باشند، مدل رگرسیون چندسطحی (آمیخته خطی) یکی از گزینه‌های مناسب برای تحلیل آماری مشاهدات است. از طرف دیگر، رگرسیون معمولی اساساً مبتنی بر مدل‌بندی میانگین شرطی متغیر پاسخ است، به این معنی که رابطه بین میانگین و متغیرهای تبیینی از طریق یک الگوی آماری نشان داده می‌شود. اگر برای مدل‌بندی یک مجموعه از داده‌ها فرض‌های اساسی مدل رگرسیون مانند نرمال بودن توزیع داده‌ها یا استقلال مولفه‌های خطا برقرار نباشد، رگرسیون مبتنی بر میانگین مناسب نخواهد بود. اگر داده‌ها شامل نقاط دوراُفتاده باشند مدل رگرسیون چندکی[۱] (QR) ، که نسبت به نقاط دوراُفتاده اُستوار بوده و توانایی ساخت الگویی برای هر نوع چندک را دارد، گزینه مناسبی برای جایگزینی رگرسیون مبتنی بر میانگین است.

پس از این‌که مدل QR توسط کاونکر و باست [۱۱] معرفی شد، تحقیقات بیشتر به قصد تعمیم حوزه کاربرد آن توسط بسیاری از محققین از جمله بوچینسکی [۱]، کاونکر و هالوگ [۱۲]، یو و همکاران [۲۲] و کاونکر [۱۰] انجام شد. از جمله، گراسی و بوتایای [۴] با در نظر گرفتن توزیع لاپلاس نامتقارن[۲] (ALD) مدل QR را از طریق درستنمایی

شرطی به چارچوب مدل‌های آمیخته تعمیم دادند. با توجه به محدودیت مدل عرض از مبدأ تصادفی برای توصیف بهتر ناهمگنی بین خوشه‌ها، گراسی و بوتایای [۵] با لحاظ اثرهای تصادفی توأم در عرض از مبدأ و شیب، مدل آمیخته چندکی خطی[۳](LQMM) را معرفی کردند. همزمان با معرفی مدل‌های متنوع QR، روش‌های برآوردیابی متفاوتی برای برآورد پارامترهای مدل گسترش یافت. از جمله آن‌ها می‌توان به روش ترکیبی تقریب مربع‌بندی گاوس[۴] (GQA) و الگوریتم بهینه‌سازی ناهموار[۵] (NSPMA) پیشنهادشده توسط گراسی و بوتایای [۵] اشاره کرد. به دلیل وجود متغیرهای پنهان اثر تصادفی در مدل LQMM، گالارزا و همکاران [۳] از تقریب تصادفی ماکسیمم‌سازی امید ریاضی[۶] (SAEM) برای پیشگویی آن‌ها استفاده کردند. این تقریب بر اساس ایده روش مونت کارلوی ماکسیمم‌سازی امید ریاضی[۷] (MCEM) که توسط وی و تنر [۲۱] پیشنهاد شده است، شکل گرفت.

در ادامه مقاله تشریح مختصری از مدل رگرسیون آمیخته چندکی خطی در بخش دوم می‌آید. سپس، جزئیاتی از برخی روش‌های برآوردیابی فرآوانی‌گرا از جمله روش ترکیبی تقریب مربع‌بندی گاوس و الگوریتم بهینه‌سازی و تقریب تصادفی ماکسیمم‌سازی امید ریاضی برای برآوردیابی پارامترهای مدل به تفکیک ارائه می‌شود. در بخش سوم، با انجام شبیه‌سازی آماری به ارزیابی عملکرد روش‌های مورد اشاره پرداخته می‌شود. بخش چهارم مقاله دربرگیرنده تحلیل یک مثال واقعی مربوط به داده‌های هزینه و درآمد خانوارهای شهری ایران در سال ۱۳۹۰ است. بحث و نتیجه‌گیری پایان بخش این مقاله است.

۲- مروری بر رگرسیون آمیخته چندکی خطی

مدل‌های اثرهای آمیخته، اغلب برای تحلیل داده‌هایی با ساختار جوامع گروه‌بندی شده مانند داده‌های طولی استفاده می‌شوند. همان‌طور که پینرو و بیتس [۱۸] اشاره کردند تمرکز اصلی چنین مدل‌هایی بر ارائه شواهدی درست و علمی از همبستگی موجود بین داده‌ها است. نکته حائز اهمیت در بکارگیری این مدل‌ها آن است که اثرهای متغیرهای تبیینی بر روی متغیر پاسخ از طریق رگرسیون مبتنی بر میانگین مدل‌بندی می‌شود. به علاوه، تبیین ساختار احتمالاتی ناهمگنی میان خوشه‌ها از طریق در نظر گرفتن توزیع نرمال برای اثرهای تصادفی و خطای مدل انجام می‌گیرد. با این حال، وقتی که شواهدی

از وجود مقادیر دور اُفتاده در بین پاسخ در اختیار باشد، چارچوب استنتاجی مبتنی بر میانگین از اعتبار لازم برخوردار نیست. مدل‌های QR ابزار بسیار مفیدی برای توصیف ارتباط رگرسیونی بین متغیرهای تبیینی و پاسخ در حضور نقاط دور اُفتاده هستند که جزئیات آن را اولین بار کاونکر و باست [۱۱] تشریح کردند.

می‌توان گفت که گراسی و بوتایای [۴] اولین محققانی بودند که تعمیمی از مدل‌های QR به مدل‌های رگرسیون آمیخته خطی را معرفی کردند. آن‌ها این کار را با درنظر گرفتن توزیع لاپلاس نامتقارن برای توزیع متغیر پاسخ در تحلیل داده‌های طولی با وجود تنها عرض از مبدأ تصادفی در مدل انجام دادند. طی سال‌های اخیر، تعمیم‌های گوناگونی از مدل گراسی و بوتایای [۴] صورت گرفت که مدل جدید پیشنهادی گراسی و بوتایای [۵] از عمومیت بیشتری برخوردار است و لذا تشریحی از آن در ادامه می‌آید.

مدل آمیخته خطی برای داده‌هایی با ساختار جامعه سلسله مراتبی به شکل

$$y_{ij} = x_{ij}^T\beta + z_{ij}^T b_i + \epsilon_{ij},\ \ i = ۱, \dots n,\ \ j = ۱, \dots, n_i \qquad (۱)$$

است که در آن y_{ij} مقدار پاسخ در اندازه‌گیری jام از یک متغیر تصادفی پیوسته برای موضوعi ام و n_i تعداد مشاهدات درون خوشه iام است. شکل ماتریسی مدل مورد اشاره را می‌توان به صورت

$$y_i = X_i\beta + Z_i b_i + \epsilon_i,\ \ i = ۱, \dots, n,$$

نوشت که در آن $y_i = (y_{i۱}, \dots, y_{in_i})^T$، $\beta = (\beta_۰, \dots, \beta_k)^T$ بردار پارامتری $(k+۱) \times ۱$، $b_i = (b_{i۱}, \dots b_{iq})^T$ بردار $q \times ۱$ اثرهای تصادفی، X_i ماتریس طرح $n_i \times (k+۱)$ شامل متغیرهای تبیینی برای خوشه‌ی iام، Z_i ماتریس طرح با بُعد $n_i \times q$ (X_i زیرماتریس Z_i) است و $\epsilon_i = (\epsilon_{i۱}, \dots, \epsilon_{in_i})^T$ خطاهای تصادفی درون خوشه‌ای هستند.

بنا به گراسی و بوتایای [۵]، تابع چندکی شرطی آمیخته خطی مدل (۱) با در نظر گرفتن ماتریس طرح Z_i با بُعد $n_i \times ۱$ (تنها یک ستون به طول n_i با عناصر یک) به صورت

$$Q_{y_{ij}|b_i}(p|x_{ij}, b_i) = x_{ij}^T\beta_p + b_i,\ \ i = ۱, \dots, n,\ \ j = ۱, \dots, n_i \qquad (۲)$$

تعریف می‌شود. بنا به لیو و بوتایای [۱۶]، روش دیگری برای بیان عبارت (۲) به صورت

$$y_{ij} = x_{ij}^T\beta_p + b_i + \epsilon_{ij},\ \ Q_{\epsilon_{ij}}(p|x_{ij}, b_i) = ۰,$$

است که در آن چندک pام خطای مدل، معروف به خطای اثر ثابت، یعنی ϵ_{ij} برابر صفر فرض می‌شود. به علاوه، فرض بر این است که خطای اثرهای ثابت ϵ_{ij} از خطای اثرهای تصادفی b_i مستقل است. همانطور که ملاحظه می‌شود ضرایب متغیرهای تبیینی X_{ij}^T تصادفی نیستند. فرض می‌شود توزیع اثر تصادفی b_i، از توزیعی مثل $f(b_i|\psi_p)$ پیروی می‌کند که در آن ψ_p ماتریس کوواریانس $q \times q$، معین مثبت و متقارن است. $\beta \in R^k$ بردار ستونی اثرهای ثابت نامعلوم با طول k و $Q_{y_{ij|b_i}}(.) \equiv F^{-۱}_{y_{ij}|b_i}(.)$ معکوس تابع توزیع تجمعی متغیر پاسخ شرطی به شرط اثر تصادفی b_i است.

مدل (۲) به LQMM با اثرهای تصادفی توأم در عرض از مبدأ و شیب با در نظر گرفتن ماتریس طرح Z_i با بُعد $n_i \times q$ به صورت

$$Q_{y_{ij}|b_i}(p|x_{ij}, b_i) = x_{ij}^T\beta_p + z_{ij}^T b_i,\ \ i = ۱, \dots, n,\ \ j = ۱, \dots, n_i, \quad (۳)$$

تعمیم داده می‌شود. مشابه حالت قبل، این عبارت را می‌توان به صورت

$$y_{ij} = x_{ij}^T\beta + z_{ij}^T b_i + \epsilon_{ij},\ \ Q_{\epsilon_{ij}}(p|x_{ij}, b_i) = ۰,$$

نیز نوشت.

قابل ذکر است که پارامترهای مدل (۳) یا هم‌ارز آن را می‌توان از طریق رویکرد فراوانی‌گرا برآورد کرد. دو روش مرسوم پیشنهادی توسط گراسی و بوتایای [۵] در این رویکرد روش ترکیبی تقریب مربع‌بندی گاوس و الگوریتم بهینه‌سازی ناهموار است. به دلیل وجود متغیرهای پنهان اثر تصادفی در مدل، گالارزا و همکاران [۳] استفاده از روش SAEM که بر پایه‌ی روش MCEM وی و تنر [۲۱] است را پیشنهاد دادند. بعد از این‌که پارامترهای LQMM به روش‌های مذکور برآورد شدند، لازم است که اثر تصادفی در مدل، پیشگویی شوند. برای پیشگویی اثر تصادفی در LQMM، گراسی و بوتایای [۵] روشی که مبتنی بر بهترین پیشگوی خطی[۸] (BLP) است را پیشنهاد دادند. از نقطه نظر استنباط آماری، انتخاب توزیع‌های مربوط به اثرهای تصادفی b_i و مشاهدات y_{ij} نقش مهمی در برآورد پارامترها و پیشگویی اثر تصادفی ایفا می‌کنند.

معمولاً فرض می‌شود، توزیع اثر تصادفی در مدل آمیخته خطی نرمال است. اما، وجود دم نازک توزیع نرمال باعث می‌شود که بعضی از نقاط دوراُفتاده توسط این توزیع به خوبی توصیف نشوند. از این رو، برای رفع این محدودیت در سال‌های اخیر توزیع‌های متفاوتی برای اثر تصادفی در نظر گرفته شد. از جمله زمانی که داده‌ها بیش از حد پراکنده باشند، بنا به لیو و بوتایای [۱۶] توزیع تی می‌تواند جایگزینی مناسب برای توزیع نرمال باشد. به‌علاوه، توزیع لاپلاس با وجود دم کلفت می‌تواند نقاط دوراُفتاده را به خوبی مدل کند. بنا به همین خاصیت، گراسی و بوتایای [۵] جایگزین اُستوارتری نسبت به توزیع نرمال برای اثرهای تصادفی پیشنهاد کردند. به زبانی دقیق‌تر، آن‌ها از توزیع لاپلاس متقارن برای مدل‌بندی اثرهای تصادفی در مدل رگرسیون آمیخته چندکی خطی استفاده کردند. شایان ذکر است که گالارزا و همکاران [۳] توزیع لاپلاس نامتقارن را برای توصیف رفتار متغیر پاسخ y_{ij} مدل (۳) در مسائل مدل‌بندی رگرسیونی پیشنهاد دادند. با پیروی از ایشان، با در نظر گرفتن اثر تصادفی مستقل $b_i \overset{i.i.d}{\sim} N_q(0, \boldsymbol{\psi})$ که در آن $\boldsymbol{\psi}$ ماتریس پراکندگی متقارن با عناصر متمایز در سلول‌هایش و $\epsilon_{ij} \sim ALD(0, \sigma, p)$، تابع چگالی شرطی توزیع لاپلاس نامتقارن $y_{ij}|b_i$ را می‌توان به صورت

$$f(y_{ij}|\beta_p, b_i, \sigma) = \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{y_{ij} - x_{ij}^T\beta_p - z_{ij}^T b_i}{\sigma}\right)\right\},$$

نوشت که در آن $\mu_{ij} = x_{ij}^T\beta_p + z_{ij}^T b_i$ پیشگویی خطی تابع چندکی pام است. لازم به ذکر است که $\rho_p(u)$ تابع زیانی معتبر بر حسب u است. بنا به گراسی [۶] در صورتی که $y_{ij} \sim ALD(\mu_{ij}, \sigma, p)$ (p پارامتر چولگی که ثابت و معلوم است) آن‌گاه تابع چندکی شرطی برابر است با $Q_{y_{ij}|b_i}(p|x_{ij}, b_i) = \mu_{ij}$. با توجه به این نکته، برای معرفی LQMM کافی است توزیع‌هایی را برای مولفه خطای مدل (ϵ_{ij}) در نظر گرفت که شرط $Q_{\epsilon_{ij}}(p|x_{ij}, b_i) = 0$ برای آن برقرار باشد.

برای برآورد پارامترهای $\theta = (\beta^T, \sigma, \psi)^T$ در LQMM، با تقریب زدن چگالی حاشیه‌ای (بخاطر پیچیدگی ساختار مولفه اثر تصادفی) تساوی

$$f(y_{ij}|\beta, b_i, \sigma) = \sigma_{n_i}^p \int_{R^q} \exp\{-\frac{1}{\sigma}\rho_p(y_i - \mu_i)\} f(b_i|\psi)\, db_i,$$

به‌دست می‌آید. با توجه به q-بُعدی بودن انتگرال اثر تصادفی، حل آن به روش مستقیم امکان‌پذیر نیست و باید از روش‌های عددی برای حل چنین انتگرالی استفاده کرد. از جمله روش‌های مرسوم برای تقریب انتگرال چندبُعدی تقریب مربع‌بندی گاوس-هرمیت و گاوس-لاگر هستند که به‌ترتیب توسط سک و دونوون [۱۹] و گلوپ و ولش [۸] معرفی شدند. لازم به اشاره است که باتوجه به بسط انتگرل روی اثر تصادفی که از توزیع نرمال پیروی می‌کند، می‌توان تقریب مربع‌بندی گاوس-هرمیت را بکار برد. به عنوان مثال چنان‌چه از روش گاوس-لاگر استفاده شود، با در نظر گرفتن توزیع لاپلاس متقارن برای اثر تصادفی، تابع چگالی حاشیه‌ای به صورت

$$f(y_i|\beta,\sigma,\psi)\simeq\sum_{k_1}^{k}\dots\sum_{k_q}^{k}\exp\left\{-\frac{1}{\sigma}\rho_p\left[y_i-X_i\beta-Z_i\psi^T\nu_{k_1,\dots,k_q}\right]\right\}\times\prod_{l=1}^{q}w_{k_l},$$

تقریب زده می‌شود که در آن k عدد صحیح، w_{k_l} وزن در نظر گرفته شده و K_jام توان چندجمله‌ای هرمیت است. به‌علاوه $\nu_{k_1,\dots,k_q}=(\nu_{k_1},\dots,\nu_{k_q})^T$ گره‌ها هستند. آنگاه لگاریتم درستنمایی تقریبی این تابع چگالی برای n گروه به شکل

$$l_{app}(\beta,\sigma,\psi|y)=\sum_i^n\log\left\{\sum_{k_1=1}^{k}\dots\sum_{k_q=1}^{k}f\left(y_i\middle|\beta,\sigma,\psi^T\nu_{k_1,\dots\nu_{k_q}}\right)\times\prod_{l=1}^{q}w_{k_l}\right\},\qquad(۴)$$

است. با به کارگیری الگوریتم بهینه‌سازی ناهموار در عبارت (۴) که از مشتق کلارک [۲] به جای مشتق‌گیری معمولی استفاده می‌کند می‌توان پارامترهای مدل را برآورد کرد. در حقیقت مشتق کلارک یک مشتق تعمیم داده شده نسبت به مشتق‌گیری معمولی است که برای بهینه‌سازی تابع‌های ناهموار بکار می‌رود. اما طبق مطالعه گالارزا و همکاران [۳] به دلیل وجود متغیرهای پنهان اثر تصادفی در مدل، می‌توان از روش دیگر که مبتنی بر استفاده از الگوریتم SAEM بر پایه الگوریتم MCEM است استفاده کرد. چون اثر تصادفی در نظر گرفته شده در رویکرد مورد اشاره متغیری پنهان است برای به دست آوردن برآورد حداکثر درستنمایی پارامترهای مدل نیاز به بهره‌گیری از الگوریتم EM است. اما

تصادفی بودن اثر، پای تقریب تصادفی را به میان می‌کشد که توجیه استفاده از الگوریتم SAEM را به همراه دارد.

بنا بر تحقیق‌های کوزوبوکسی و پوجوسکی [۱۴] و کوتز و همکاران [۱۳]، با فرض $U\sim\exp(\sigma)$ و $Z\sim N(0,1)$ که دو متغیر تصادفی مستقل از هم هستند، می‌توان Y را به‌صورت ترکیب خطی به شکل

$$Y=\mu+\vartheta_p U+\tau_p\sqrt{\sigma U}Z, \qquad (۵)$$

نوشت که در آن صورت $Y\sim ALD(\mu,\sigma,p)$، $\vartheta_p=\frac{1-2p}{p(1-p)}$ ، $\tau_p^2=\frac{2}{p(1-p)}$ و اندیس p نشان دهنده پارامتر چولگی هستند. نوشتن LQMM به صورت تساوی (۵) نه تنها روندی برای شبیه‌سازی از آن را تسهیل می‌کند بلکه چارچوبی برای نگارش آن به صورت سلسله مراتبی (استنباط از طریق چگالی‌های شرطی) فراهم می‌نماید. به طور دقیق‌تر، برای برآورد پارامترهای LQMM با استفاده از الگوریتم SAEM، ساختار سلسله مراتبی ALD معادله (۵) به صورت

$$y_i|b_i,u_i\sim N_{n_i}\left(X_i\beta_p+Z_ib_i+\vartheta_p u_i,\sigma\tau_p^2 D_i\right),$$

$$b_i\sim N_q(0,\psi),$$

$$u_i\sim\prod_{j=1}^{n_i}Exp(\sigma),$$

در نظر گرفته می‌شود که در آن D_i ماتریس قطری و شامل بردار مقادیر پنهان $u_i=(u_{i1},\dots,u_{in_i})^T$ است. اکنون برای به دست آوردن برآورد پارامترهای مدل، می‌توان مراحل الگوریتم EM را دنبال کرد. لازم به ذکر است که برآورد پارامترها در مرحله kام الگوریتم با نماد $\theta^{(k)}=(\beta_p^{(k)T},\sigma^k,\alpha^{(k)T})^T$ مشخص می‌شود. معمولا در مرحله اول اجرای الگوریتم مقادیری از پیش داده شده برای این بردار نظر گرفته می‌شود. با استفاده از لگاریتم تابع درستنمایی مشاهدات کامل به شرط داده‌های مشاهده شده که از ضرب چگالی‌های نرمال و نمایی اشاره شده در بالا با لحاظ تعداد مولفه‌های درگیر

اندیس‌های i و j به دست می‌آید، کمیت امید ریاضی شرطی که آن را با تابع $Q(\theta|\hat{\theta}^{(k)})$ نشان می‌دهیم، به صورت

$$Q(\theta|\hat{\theta}^{(k)}) = \sum_{i=1}^{n} Q_i(\theta|\hat{\theta}^{(k)}),$$

نوشته می‌شود که در آن

$$\begin{aligned} Q_i(\theta|\hat{\theta}^{(k)}) &= E\{\ell_c(\theta; y_{i_c})|\theta^{(k)}, y_i\} \\ &= -\frac{3}{2} n_i \log\sigma - \frac{1}{2\sigma\tau_p^2}[(y_i - X_i\beta_p)^T \widehat{D_i^{-1}}^{(k)}(y_i - X_i\beta_p) \\ &\quad -2(y_i - X_i\beta_p)\left(\widehat{D_i^{-1}Zb}\right)_i^{(k)} + tr\left\{Z_i\left(bb^T\widehat{ZD_i^{-1}}\right)_i^{(k)}\right\} \\ &\quad -2\vartheta_p(y_i - X_i\beta_p)^T 1_{n_i} + 2\vartheta_p\left(Z\hat{b}^k\right)_i^{(k)} 1_{n_i} + \frac{\tau_p^4}{4}\widehat{u_i}^{(k)T} 1_{n_i} \\ &\quad -\frac{1}{2}\log|\psi| - \frac{1}{2} tr\left\{\widehat{bb^T}_i^{(k)} \psi^{-1}\right\}. \end{aligned} \qquad (۶)$$

در این مرحله از الگوریتم باید امید ریاضی‌های شرطی مربوطه در(۶) -عباراتی با نماد برآورد روی آن‌ها- محاسبه شوند. با توجه به این‌که توزیع توأم متغیرهای پنهان $(b_i^{(k)}, u_i^{(k)})$ نامعلوم است، نمی‌توان امید ریاضی‌های شرطی را به صورت تحلیلی محاسبه کرد. از این رو، روشی برای تقریب امید ریاضی شرطی توسط گالارزا و همکاران [۳] پیشنهاد شد. به این روش، الگوریتم MCEM گفته می‌شود. در فرآیند الگوریتم مورد اشاره، با استفاده از الگوریتم‌های شبیه‌سازی مونت کارلوی زنجیر مارکوفی همانند متروپولیس-هستینگس[۹] (MH) (متروپولیس و همکاران [۱۷]؛ هستینگس [۹]) از چگالی حاشیه‌ای شرطی $f(b_i|\theta^{(k)}, y_i)$ می‌توان نمونه‌های $b_i^{(\ell,k)}$ تولید کرد و امید ریاضی متغیر پنهان u_i را براساس چگالی شرطی کامل $f(u_i|y_i, b_i)$ به دست آورد. قابل ذکر است که در الگوریتم MCEM، تعداد شبیه‌سازی مونت کارلو تا حدی زیاد است. به‌علاوه، این تعداد شبیه‌سازی مستقل مربوط به داده‌های پنهان برای رسیدن به یک

تقریب خوب از پارامترها، محاسبات سنگینی را به دنبال دارد. از این رو، از الگوریتم SAEM نسبت به الگوریتم MCEM که تعداد شبیه‌سازی‌های مستقل داده‌های پنهان از یک تابع چگالی شرطی در هر تکرار الگوریتم، کمتر یا مساوی ۲۰ و امید ریاضی شرطی آن که بر اساس یک تقریب تصادفی استوار است، استفاده می‌شود.

فرایند الگوریتم مذکور در تکرار k-ام به این صورت است که در مرحله اول به ازای $\ell = ۱, \dots, m$ که $m \leq ۲۰$، مقادیر پنهان $b_i^{(\ell,k)}$ با استفاده از روش‌های شبیه‌سازی از تابع چگالی شرطی$f(b_i|\theta^{(k)}, y_i)$ ، تولید می‌شود. در همین مرحله با استفاده از تقریب تصادفی، تابع $Q\left(\theta\middle|\hat{\theta}^{(k)}\right)$ به صورت

$$Q\left(\theta\middle|\hat{\theta}^{(k)}\right) \approx Q\left(\theta\middle|\hat{\theta}^{(k-1)}\right) + \delta_k \left[\frac{۱}{m}\sum_{\ell=۱}^{m} \ell_c\left(\theta, y_{obs}, b_i^{(\ell,k)}\middle|\theta^{(k)}, y_{obs}\right) - Q\left(\theta\middle|\hat{\theta}^{(k-۱)}\right)\right],$$

به روزرسانی می‌شود که در آن δ_k بنا به کوهن و لویل [۱۵] پارامتر هموار است. در مرحله دوم، پارامترهای $\hat{\theta}^k = (\hat{\theta}_p^k, \hat{\sigma}^k, \hat{\psi}^k)$ با ماکسیمم کردن تابع $Q\left(\theta\middle|\hat{\theta}^{(k)}\right)$، به روزرسانی می‌شوند.

۳- آزمایش شبیه‌سازی

هدف این بخش مقایسه عملکرد روش ترکیبی تقریب مربع‌بندی گاوس، الگوریتم بهینه‌سازی ناهموار و الگوریتم SAEM در برآورد پارامترها در LQMM بر اساس انجام شبیه‌سازی آماری است. از این رو، برای ارزیابی رویکردهای مذکور از معیارهایی شامل اریبی نسبی[۱۰](RBias) ، انحراف استاندارد مونت کارلو[۱۱](MC-Sd) و ریشه میانگین توان دوم خطا [۱۲] (RMSE) استفاده می‌شود. لازم به اشاره است که کدهای مورد نیاز و تحلیل‌های آماری مرتبط با آن‌ها در این بخش و بخش بعدی، همگی در نرم افزار R نسخه ۳.۱.۳ اجرا شده‌اند.

مدل رگرسیون آمیخته خطی

$$y_i = X_i\beta + Z_i b_i + \epsilon_i,\ \ i = ۱, \dots, n, \quad (۷)$$

را در نظر بگیرید. برای سهولت نمادگذاری، تعداد گروه‌ها و اندازه گروه‌ها به ترتیب با نماد NG[۱۳] و GS[۱۴] نشان داده می‌شود. هم‌چنین، ماتریس طرح Z_i به‌گونه‌ای در نظر گرفته شد که منجر به وجود دو متغیر کمکی در سطح فردی و یک متغیر کمکی در سطح گروهی شود.

بر اساس ۱۰۰= M بار تکرار در فرایند برازش مدل در بخش ۳-۱ که در ادامه می‌آید (تعداد حجم نمونه در بخش بعد مشخص می‌شود) معیارهای

$$RBias = \frac{|Bias(\hat{\theta}_\ell)|}{\theta_\ell},$$

و

$$MC - Sd(\hat{\theta}_\ell) = \sqrt{\frac{۱}{M-۱}\sum_{j=۱}^{M}\left(\hat{\theta}_\ell{}^{(j)} - \bar{\hat{\theta}}_\ell\right)^۲}$$

محاسبه شدند که در آن $Bias(\hat{\theta}_\ell) = \bar{\hat{\theta}}_\ell - \theta_\ell$ و

$$RMSE(\hat{\theta}_\ell) = \sqrt{[MC - Sd(\hat{\theta}_\ell)]^۲ + [Bias(\hat{\theta}_\ell)]^۲}.$$

به علاوه، میانگین مونت کارلو به صورت $\bar{\hat{\theta}}_\ell = \frac{۱}{M}\sum_{j=۱}^{M}\hat{\theta}_\ell{}^{(j)}$ در نظر گرفته شده که $\theta_\ell^{(j)}$ برآورد θ_ℓ از نمونه‌ی jام به ازای $j = ۱, \dots, M$ است.

در این قسمت، با انجام شبیه‌سازی عملکرد پارامترهای برآورد شده در LQMM با استفاده از الگوریتم SAEM مطرح شده توسط گالارزا و همکاران [۳] و روش تقریبی پیشنهاد شده توسط گراسی و بوتایای [۵] با چندک‌های مختلف $p = \{۰/۲۵, ۰/۷۵\}$، مورد مقایسه قرار می‌گیرد. برای این امر، نمونه‌ای از مدل رگرسیونی (۷) تولید شدند که در آن ستون‌های ماتریس طرح X_i و Z_i به غیر از ستون اول، با تولید نمونه از $N_۲(۰, I_۲)$ شکل می‌گیرند. به علاوه خطای مدل از $ALD(0, \sigma, p)$ و اثر تصادفی از $N_۲(۰, \psi_{۲\times۲})$ تولید شدند. پارامترهای اثرهای ثابت به صورت‌های $\beta_۰ = ۰/۸$ ، $\beta_۱ = ۰/۵$ ، $\beta_۲ = ۱$ و

σ_ϵ =۰/۲ و ماتریس $\psi_{۲\times۲}$ با عناصر $\psi_{۱۱}$ =۰/۸ ، $\psi_{۱۲}$ =۰/۵ ، $\psi_{۲۲}$ =۱ در نظر گرفته شد. برای مدل مورد اشاره، ۱۰۰ مجموعه داده با NG برابر ۱۰۰، ۲۰۰ و ۳۰۰ و GS برابر ۱۰، شبیه‌سازی و برای هر مجموعه داده، پارامترهای مدل بر اساس چندک‌های ۲۵ام و ۷۵ام برآورد شدند. مقایسه روش‌های برآوردیابی بر اساس معیارهای تشریح شده در جدول ۱ آمده است. توجه به این نکته حائز اهمیت است که برای اجرای الگوریتم SAEM، از کتابخانه qrLMM و الگوریتم ارائه شده توسط گراسی و بوتایای [۵] از کتابخانه lqmm استفاده شده است.

طبق جدول ۱ مشاهده می‌شود که مقادیر معیار RMSE برای پارامترهای برآورد شده $\hat{\beta}_۰$ و $\hat{\beta}_۱$ با در نظر گرفتن چندک‌ها و تعداد گروه‌ها با استفاده از روش برآوردیابی گراسی و بوتایای [۵]، که در جدول مورد اشاره با کلمه اختصار GB مشخص شد، با اختلاف نسبتا بالایی کمتر از الگوریتم SAEM است. به علاوه، با افزایش تعداد گروه‌ها برای هر چندک، مقادیر معیار RMSE برای روش گراسی و بوتایای نسبت به الگوریتم SAEM به شکل یک روند نزولی است. اما از لحاظ پایین بودن مقادیر معیار مربوطه و وجود روند نزولی آن‌ها در حالت‌های خاصی، چنین امری برای پارامترهای $\hat{\beta}_۲$ و $\hat{\sigma}_\epsilon$ محقق نشده است. این امر ممکن است ناشی از کم در نظر گرفتن تعداد نمونه‌ها داخل گروه‌ها و عدم تصادفی بودن ضریب رگرسیونی $\beta_۲$ باشد. نتیجه کلی حاصل از انجام مطالعه شبیه‌سازی این است که روش تقریبی گراسی و بوتایای [۵] برآوردهای دقیق‌تری نسبت به الگوریتم SAEM پیشنهاد شده توسط گالارزا و همکاران [۳] ارائه می‌دهد.

۴- برازش مدل رگرسیون آمیخته چندکی خطی برای تحلیل داده‌های هزینه و درآمد خانوار

داده‌های هزینه و درآمد خانوارهای شهری مربوط به سال۱۳۹۰، دربرگیرنده این اطلاعات برای ۱۸۷۲۷ خانوار از ۳۱ استان کشور است. به دلیل محدودیت مدل رگرسیون آمیخته در اندازه نمونه در سطوح بالاتر، انتخاب استان‌ها از بین شش منطقه متفاوت جغرافیای

جدول ۱- مقایسه ارزیابی عملکرد مدل آمیخته چندکی خطی با اندازه نمونه ده‌تایی برای هر گروه.

RMSE									
$\widehat{\sigma_\epsilon}$		$\widehat{\beta_۲}$		$\widehat{\beta_۱}$		$\widehat{\beta_۰}$		پارامتر	
GB	SAEM	GB	SAEM	GB	SAEM	GB	SAEM	NG	چندک
۰/۰۰۴۹	۰/۰۰۱۲	۰/۰۰۷۳	۰/۰۰۴۸	۰/۱۰۹۴	۰/۱۹۵	۰/۱۷۹۴	۰/۳۳۵۸	۵۰	
۰/۰۰۶۷	۰/۰۰۳۳	۰/۰۰۳۲	۰/۰۰۳۵	۰/۰۷۵۳	۰/۱۵۹۸	۰/۱۳۳۲	۰/۳۳۳۳	۱۰۰	۲۵
۰/۰۰۷۵	۰/۰۰۲۵	۰/۰۰۱۵	۰/۰۰۲۹	۰/۰۷۱۴	۰/۱۵۶۷	۰/۱۰۲۲	۰/۳۳۷۹	۲۰۰	
۰/۰۰۵	۰/۰۰۰۶	۰/۰۱۰۸	۰/۰۰۱۹	۰/۱۱۱۱	۰/۱۶۸۵	۰/۱۵۷	۰/۳۲۸۵	۵۰	
۰/۰۰۷	۰/۰۰۱۵	۰/۰۰۲۶	۰/۰۰۲	۰/۰۹۶۷	۰/۱۷۴	۰/۱۳۳۹	۰/۳۲۸۹	۱۰۰	۷۵
۰/۰۰۷۵	۰/۰۰۲۸	۰/۰۰۳۷	۰/۰۰۰۸	۰/۰۵۶۹	۰/۱۶۴۱	۰/۰۹۶۸	۰/۳۳۷۳	۲۰۰	

کشور انجام گرفت. اسامی استان‌ها که در جدول ۳ آمده است موید این مطلب است. حجم نمونه برای استان‌ها بر اساس تخصیص متناسب با جمعیت استان‌های انتخاب شده تعیین شد. هم‌چنین برای انتخاب نمونه (خانوار) در هر استان از نمونه‌گیری تصادفی ساده استفاده شد. در نهایت، تعداد خانوارهای مورد بررسی در این مطالعه، ۲۰۰۰ خانوار شهری از ۱۸ استان بوده است. به منظور تحلیل داده‌های مورد اشاره، متغیر پاسخ هزینه ناخالص خانوارها (y) و متغیرهای تبیینی درآمد خالص ($X۱$)، سن سرپرست خانوار ($X۲$)، سطح زیربنای محل سکونت ($X۳$)، بُعد خانوار ($X۴$)، تعداد افراد شاغل ($X۵$)، تعداد افراد با درآمد ($X۶$)، جمعیت استان ($X۷$)، جنسیت ($X۸$)، وضع فعالیت سرپرست خانوار ($X۹$)، وضع زناشویی سرپرست خانوار ($X۱۰$) و نحوه تصرف محل سکونت ($X۱۱$) در نظر گرفته شد. اما، متغیرهای معنادار در مدل نهایی گزارش شد. با لحاظ اثر تصادفی استان‌ها ،(b_i) مدل رگرسیون آمیخته چندکی خطی به صورت

$$y_{ij} = \beta_{۰j} + \beta_{۱j} x1_{ij} + \beta_{۲j} x3_{ij} + \beta_{۳j} x2_{ij} + \beta_{۴} x۴_{ij} + \beta_{۵} x۵_{ij} + \beta_{۶} x۷_{ij} + \sum_{k=۱}^{۵} \beta_{۷}^{(k)} x۹_{i}^{(k)} \quad (۸)$$

$$+\sum_{k=1}^{4}\beta_{8}^{(k)}x11_{i}^{(k)}+\varepsilon_{ij}$$
$$i=1,\dots,18,\ \ j=1,\dots,n_i$$
$$\beta_{0i}=\beta_{0}+b_{0i}$$
$$\beta_{1i}=\beta_{1}+b_{1i}$$
$$\beta_{2i}=\beta_{2}+b_{2i}$$

مدنظر قرار گرفته شد که در آن چندک pام، خطای اثرهای ثابت ϵ_{ij} برابر صفر در نظر گرفته شده است. لازم به ذکر است که برای آن دسته از متغیرهای کیفی که دارای برچسب‌های متنوع هستند از نام متغیر متناظر همراه با بالانویس k استفاده شد. به‌علاوه در این مدل، متغیرهای تبیینی کیفی که بیش از دو سطح دارند از جمله وضع فعالیت و نحوه تصرف محل سکونت به عنوان فاکتور در نظر گرفته و بر اساس تلفیق مناسب به صورت صفر و یک به متغیرهای دودویی تبدیل شدند. مثلا مالکین و غیرمالکین به ترتیب با برچسب‌های یک و صفر مشخص شدند. سپس، رگرسیون آمیخته چندکی خطی (۸) با در نظر گرفتن دو چندک متفاوت $p=\{0.14, 0.75\}$ و توزیع ALD در نرم افزار R نسخه ۳.۱.۳ با استفاده از بسته lqmm برازش داده شد. خلاصه نتایج به صورت جدول ۲ گزارش شد. به علاوه، اگر چه نتایج حاصل از انتخاب چندک ۱۴ام مشابه مقادیر متناظر از چندک ۲۵ام بود ولی تفسیرپذیری بهتر اولی ما را متقاعد به گزارش آن کرده است. لازم به اشاره است معیار انتخاب چندک‌های مذکور براساس پارامتر چولگی در نظر گرفته شده در توزیع لاپلاس نامتقارن هنگام برازش توزیع اشاره شده به داده‌های هزینه و درآمد، که نشان از پیروی توزیع داده‌ها از توزیع مورد نظر داشت، بود. توجه شود که نتایج خروجی تحلیل شامل مقادیر p−مقدارها و فواصل اطمینان تقریبی برای پارامترها بود اما برای جلوگیری از حجیم شدن مقاله از ارائه آن‌ها خودداری شده است.

بر اساس مدل (۸)، در بررسی تغییرات هزینه ناخالص خانوارها علاوه بر تصادفی درنظر گرفتن استان محل سکونت افراد، اثرهای درآمد و سطح زیربنا (به صورت توام) درون هر استان تصادفی را در نظر گرفتیم و در نتیجه اثرهای تصادفی b_{1i} و b_{2i} را نیز به این متغیرها القا کردیم. با توجه به غیر قابل مشاهده بودن اثرهای تصادفی با استفاده از روش

جدول ۲- برازش مدل (۸) به روش گراسی و بوتایای (۲۰۱۴) با چندک‌های متفاوت ۱۴ام و ۷۵ام

چندک	پارامتر	Geraci	پارامتر	Geraci
	β_0	-۰/۰۰۷	β_5	-۰/۱۴
	β_1	۰/۶۵۵	β_6	۰/۰۷۲
۱۴	β_2	۰/۱۵	β_7	۰/۰۸۵
	β_3	-۰/۰۰۸	β_8	۰/۰۶
	β_4	۰/۰۰۰۰۲	σ_ϵ	۰/۱۱
	β_0	۰/۰۴۸	β_5	-۰/۰۹۲
	β_1	۰/۷۳	β_6	۰/۰۵۳
۷۵	β_2	۰/۱۶	β_7	۰/۱۴۴
	β_3	-۰/۰۰۰۴	β_8	۰/۰۷۲
	β_4	۰/۰۶۵	σ_ϵ	۰/۱۹۲

گراسی و بوتایای (۲۰۱۴) پیشگویی آن‌ها انجام شد و طبق جدول ۳، پارامترهای β_{0i}، β_{1i} و β_{2i} برآورد شدند. طبق جدول ۲ بر اساس چندک ۱۴ام، هزینه ناخالص خانوارهای در نظر گرفته شده در استان‌های مختلف می‌توان به تفسیر نتایج به دست آمده از مدل رگرسیون آمیخته چندکی خطی پرداخت. با توجه به پارامترهای برآورد شده در مدل (۸) بر اساس داده‌های هزینه و درآمد ۱۳۹۰، می‌توان گفت که انتظار می‌رود درآمد خالص خانوارها در استان‌ها، با فرض تغییر نکردن تمامی عوامل در خانوارها، تقریبا به اندازه‌ی ۶۵% بر هزینه‌ی ناخالص خانوارها در استان‌های مختلف تأثیر می‌گذارد. افزایش هزینه برای خانوارهای پر جمعیت و استان‌هایی با جمعیت فراوان چشم‌گیر است. هزینه ناخالص خانوارهایی که سرپرست آن‌ها شاغل هست و یا اینکه خانوار اجاره‌نشین هستند بیشتر خواهد بود. از طرفی، افزایش سن سرپرست خانوار و داشتن تعداد افراد شاغل بیشتر در یک خانواده، منجر به کاهش هزینه آن خانواده می‌شود. طبق جدول۳، با در نظر گرفتن اثر استان، انتظار می‌رود هزینه ناخالص یک خانوار در استان چهارمحال و بختیاری که درآمد آن در مقایسه با خانوار دیگری در همین استان به اندازه یک میلیون ریال بیشتر است، به اندازه ۸۱۲۰۰۰ ریال بیشتر باشد. لازم به ذکر است مساحت

زیربنای محل سکونت نمی‌تواند تأثیر چندان زیادی در هزینه ناخالص خانوارها در هر یک از استان‌ها داشته باشد.

طبق جدول ۲ با در نظر گفتن چندک ۷۵ام هزینه ناخالص خانوارها انتظار می‌رود درآمد با فرض ثابت نگه داشتن بقیه عوامل به اندازه‌ی ۷۳% در هزینه ناخالص تأثیر بگذارد. به علاوه، هزینه برای خانوارهای با تعداد بالاتر و افزایش جمعیت استان‌ها افزایش می‌یابد. انتظار می‌رود این افزایش در خانوارهای با سرپرست شاغل و اجاره‌نشین بودن آن‌ها محقق شود. اما با افزایش سن سرپرست خانوارها و تعداد افراد شاغل در خانوارها انتظار کاهش هزینه آن‌ها را داریم. طبق جدول ۳، با در نظر گرفتن اثر استان، درآمد می‌تواند تأثیر قابل ملاحظه‌ای در هزینه ناخالص خانوارها در هر یک استان‌ها داشته باشد. به عنوان مثال، اگر تفاوت درآمد خالص یک خانوار در استانی مانند گیلان نسبت به خانواری دیگر در استان مربوطه با فرض یکسان بودن تمامی عوامل به اندازه یک میلیون ریال باشد، هزینه ناخالص در میان دو خانوار به اندازه ۸۲۶۰۰۰ ریال تفاوت دارد. انتظار می‌رود مساحت بالای سطح زیربنای محل سکونت خانوارها در هر یک از استان‌ها منجر به افزایش هزینه ناخالص آن خانواده شود.

بنا به تحلیل انجام گرفته بر پایه مدل پیشنهادی و نتایج حاصل از آن برای داده‌های هزینه و درآمد با چندک‌های ۱۴ام و ۷۵ام، می‌توان اظهار داشت که درآمد خالص خانوارها در استان‌های مختلف تأثیر بسزایی در هزینه ناخالص خانوارها دارد. این بدان معنی است که به طور کلی هزینه در استان‌ها در ارتباط تنگاتنگ با میزان درآمد مردمان همان استان است. اما، در نگاه کلی به کشور ایران می‌توان گفت که مساحت زیربنای خانه‌ها، تعداد افراد هر خانوار، جمعیت کلی استان، شاغل بودن سرپرست خانوار و وضعیت مالکیت خانوارها (به ویژه اجاره‌نشینی) تاثیر معنی‌داری بر هزینه ناخالص خانوارها ندارد. نکته قابل تأمل این است که افزایش سن سرپرست خانوار و تعداد افراد شاغل در هر خانوار تأثیر معکوسی در هزینه ناخالص خانوارها دارد که این نکته با توجه به بافت اجتماعی، اقتصادی و فرهنگی ایران قابل توجیه است.

جدول ۳- برآورد ضرایب متغییرهای تبیینی برای چندک ۱۴ام و ۷۵ام

چندک	۱۴			۷۵		
نام استان	$\hat{\beta}_{\circ i}$	$\hat{\beta}_{۱i}$	$\hat{\beta}_{۲i}$	$\hat{\beta}_{\circ i}$	$\hat{\beta}_{۱i}$	$\hat{\beta}_{۲i}$
کرمان	-۰/۲۱۵	۰/۵۸۷	۰/۰۸۶	۰/۰۵۷	۰/۵۹۷	۰/۱۰۲
فارس	-۰/۰۵۳	۰/۵۸۶	۰/۱۲۳	۰/۲۰۶	۰/۶۱۳	۰/۱۶۸
هرمزگان	-۰/۰۲۲	۰/۶۳۶	۰/۱۷	۰/۲۵	۰/۶۷۱	۰/۱۹۵
گلستان	-۰/۱۳۵	۰/۷۰۹	۰/۱۵۶	۰/۱۲۳	۰/۷۰۴	۰/۱۸۲
خراسان رضوی	-۰/۲۳۴	۰/۶۴۱	۰/۰۷۲	۰/۰۳۱	۰/۶۷	۰/۱
سمنان	-۰/۱۴۶	۰/۶۲۷	۰/۱۱	۰/۱۱۵	۰/۷۳۴	۰/۱۹۶
همدان	-۰/۱۶۳	۰/۷۲۲	۰/۱۳۸	۰/۰۹۹	۰/۷۳	۰/۱۷
لرستان	۰/۱۲۳	۰/۷۲۷	۰/۲۲۷	۰/۳۵۶	۰/۷۳۷	۰/۲۷
کرمانشاه	۰/۰۱۳	۰/۶۷۵	۰/۱۸۲	۰/۲۷۱	۰/۶۸	۰/۲۱۲
گیلان	۰/۰۷۳	۰/۷۹۵	۰/۲۹۸	۰/۳۱۸	۰/۸۲۶	۰/۳۱۸
اردبیل	۰/۰۲۹	۰/۷۳۹	۰/۲۱۴	۰/۲۹۱	۰/۷۵۶	۰/۲۵
کردستان	-۰/۰۵۳	۰/۶۸	۰/۱۵۲	۰/۱۵۲	۰/۶۷۸	۰/۱۵۴
یزد	-۰/۳۰۸	۰/۵۴۳	۰/۰۲۲	-۰/۰۴۵	۰/۵۹۶	۰/۰۸
اصفهان	۰/۰۶۸	۰/۷۶۹	۰/۲۰۲	۰/۳۶۵	۰/۷۷	۰/۲۶۳
چهارمحال و بختیاری	۰/۰۹۷	۰/۸۱۲	۰/۲۱۴	۰/۳۲۵	۰/۸۵۹	۰/۲۹۲
تهران	-۰/۰۸۴	۰/۷۵۴	۰/۲۱۶	۰/۲۹۹	۰/۷۶۵	۰/۲۲۷
قم	-۰/۱۶۵	۰/۶۷۷	۰/۱۱۸	۰/۰۸۶	۰/۶۸۷	۰/۱۴۸
مازندران	-۰/۱۱۱	۰/۶۳۵	۰/۱۱	۰/۱۳۹	۰/۶۸۲	۰/۱۶

۵- بحث و نتیجه‌گیری

یکی از فرض‌های رایج مدل‌های رگرسیون خطی، ناهمبسته بودن خطاهای مدل است. با این حال، در صورت نقض این فرض و به طور دقیق‌تر همبستگی درون گروهی بین مشاهدات، می‌توان از مدل‌های رگرسیون آمیخته خطی استفاده کرد. معمولاً ساختار چنین مدل‌هایی نیز بر اساس انحراف مشاهدات از میانگین و به علاوه فرض توزیع نرمال برای داده‌ها بنا شده است. برخی اوقات، محققین علاقمندند از این دو محدودیت عبور کرده و مدل‌های اُستوارتری را به خدمت بگیرند که چنین امکانی توسط مدل‌های رگرسیون آمیخته چندکی فراهم می‌شود. در این مقاله، مدل رگرسیون آمیخته چندکی در تحلیل داده‌های طرح هزینه و درآمد خانوارهای شهری ایرانی در ۱۳۹۰ به کار گرفته شد. در

مطالعه این مثال واقعی، هدف، بررسی عوامل موثر بر هزینه ناخالص خانوارها در ایران بوده است. نتایج حاصل شده نشان می‌دهد که با افزایش درآمد خانوار، هزینه ناخالص آن‌ها افزایش چشمگیری می‌تواند داشته باشد. این موضوع نشان می‌دهد که درآمد به عنوان متغیر تبیینی، تأثیر چشمگیری در هزینه ناخالص خانوارها در استان‌های مختلف دارد. از طرفی سطح زیربنای محل سکونت، بُعد خانوار، جمعیت استان، وضع فعالیت سرپرست خانوار و نحوه تصرف محل سکونت خانوارها تأثیر کمتری در هزینه ناخالص خانوارها دارند. به‌علاوه، با افزایش سن سرپرست خانوار و وجود تعداد افراد شاغل بیشتر در خانوارها، هزینه ناخالص سرپرست خانوار کاهش پیدا می‌کند. معمولا، ایرانی‌های با درآمد بالاتر سعی در صرف هزینه بیشتر دارند چرا که نگران زندگی مرفه‌تری هستند. همچنین، وجود تعداد افراد بیشتر در خانوارها، دغدغه بیشتر سرپرست خانوار را در پی دارد و این منجر به تحمیل هزینه بیشتر می‌شود. از طرفی دیگر، علاوه بر داشتن شغل سرپرست خانوار شاغل بودن اعضای دیگر خانواده کاهش هزینه صرف شده توسط سرپرست را به همراه دارد.

توضیحات

1. Quantile Regression
2. Asymmetric Laplace Distribution
3. Linear Quantile Mixed Model
4. Gaussian Quadrature Approximations
5. Non-Smooth Optimization Algorithms
6. Stochastic Approximation of the EM
7. Monte Carlo EM
8. Best Linear Predictor
9. Metropolis-Hastings
10. Relative Bias
11. Monte Carlo Standard Deviation
12. Root Mean Squared Error
13. Number of Groups
14. Groups Size

مرجع‌ها

[1] Buchinsky, M. (1998). Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research, Journal of Human Resources, **33**, 88–126.

[2] Clarke, F.H. (1990). Optimization and Nonsmooth Analysis, SIAM, Philadelphia.

[3] Galarzaa, C.E., Bandyopadhyayb, D. and Lachosa, V.H. (2017). Quantile Regression for Linear Mixed Models: A Stochastic Approximation EM Approach, Statistics and Interface, **10**, 471–482.

[4] Geraci, M. and Bottai, M. (2007). Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution, Biostatistics, **8**, 140–154.

[5] Geraci, M. and Bottai, M. (2014). Linear Quantile Mixed Models, Statistics and Computing, **24**, 461–479.

[6] Geraci, M. (2011). lqmm: Estimating Quantile Regression Models for Independent and Hierarchical Data with R, R Users Conference, University of Warwick, Warwick, UK.

[7] Goldstein, H. (2011). Multilevel Statistical Models, John Wiley and Sons, Chichester.

[8] Golub, G.H. and Welsch, J.H. (1969). Calculation of Gaussian Quadrature Rules, Mathematics of Computation, **23**, 221–230.

[9] Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications, Biometrika, **57**, 97–109.

[10] Koenker, R. (2005). Quantile Regression, Cambridge University Press, New York.

[11] Koenker, R. and Bassett, G. (1978). Regression Quantiles,

Econometrica, **46**, 33–50.

[12] Koenker, R. and Hallock, K. (2001). Quantile Regression: An introduction, Journal of Economic Perspectives, **15**, 43–56.

[13] Kotz, S., Kozubowski, T. and Podgorski, K. (2012). The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Economics, Engineering and Finance, Springer Science and Business Media, New York.

[14] Kozubowski, T.J. and Podgorski, K. (2000). A Multivariate and Asymmetric Generalization of Laplace Distribution, Computational Statistics, **15**, 531–540.

[15] Kuhn, E. and Lavielle, M. (2004). Coupling a Stochastic Approximation Version of EM with an MCMC Procedure, ESAIM: Probability and Statistics, **8**, 115–131.

[16] Liu, Y. and Bottai, M. (2009). Mixed-Effects Models for Conditional Quantiles with Longitudinal Data, The International Journal of Biostatistics, **5**, 1–22.

[17] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, E. (1953). Equation of State Calculations by Fast Computing Machines, The Journal of Chemical Physics, **21**, 1087–1092.

[18] Pinheiro, J.C. and Bates, D.M. (2000). Mixed-Effects Models in S and S-PLUS, Springer, New York.

[19] Sack, R.A. and Donovan, A.F. (1971). An Algorithm for Gaussian Quadrature Given Modified Moments, Numerische Mathematik, **18**, 465–478.

[20] Snijders, T. and Bosker, R. (2000). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, Sage Publications, London.

[21] Wei, G.C. and Tanner, M.A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms, Journal of the American statistical Association, **85**, 699–704.

[22] Yu, K., Lu, Z. and Stander, J. (2003). Quantile Regression: Applications and Current Research Areas, Journal of the Royal Statistical Society, **52**, 331–350.

عنایت بارانی
فوق لیسانس آمار
تهران، بزرگراه جلال آل احمد، دانشگاه تربیت مدرس، گروه آمار.
رایانشانی: enayat.barani@gmail.com

موسی گل‌علیزاده
دکتری آمار
تهران، بزرگراه جلال آل احمد، دانشگاه تربیت مدرس، گروه آمار.
رایانشانی: golalizadeh@modares.ac.ir