مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۱، شماره‌ی ۱، بهار و تابستان ۱۳۸۹، صص ۸۹– ۱۰۸

مقایسه‌ی جانهی الگوریتم EM با دو روش جانهی میانگینی و نمونه‌های جدید در آمارگیری‌های پانلی

آسیه رشیدی‌نژاد* و حمیدرضا نواب‌پور

دانشگاه علامه‌طباطبایی

چکیده: در اقتصاد و سایر علوم اجتماعی، پژوهش‌گران اغلب تمایل به مدل‌بندی داده‌های پانلی که در آن واحدهای نمونه‌ای به‌طور مکرر در مقاطع زمانی مختلف مشاهده می‌شوند، دارند. یکی از کاربردهای داده‌های پانلی براورد نرخ تغییر میانگین متغیر پاسخ در طی زمان است. در تمام آمارگیری‌ها به ویژه آمارگیری‌های پانلی، بی‌پاسخی یک مشکل اساسی است که در داده‌های علوم اجتماعی و پزشکی به وفور رخ می‌دهد. این نوع مطالعه‌ها معمولاً با کاهش پاسخگو در دوره‌های دوم به بعد تولید داده‌ها مواجه هستند. این امر که منجر به نمونه‌ی کاهیده می‌شود سبب کاهش کارایی براوردگرها و غالباً نیز سبب اریبی آن‌ها می‌شود. برای برخورد با این مشکل در آمارگیری پانلی روش‌های «جانهی» و «وزن‌دهی» گوناگونی وجود دارد که از جمله‌ی این روش‌ها، جانهی با الگوریتم EM (Expectation Maximization Algorithm) می‌باشد. در این مقاله پس از معرفی مفهوم‌های اولیه آمارگیری پانلی، انواع گم‌شدگی در آمارگیری‌های پانلی و ساختارهای گم‌شدگی، الگوریتم EM به‌عنوان روشی برای جانهی داده‌های گم‌شده معرفی می‌شود. سپس با استفاده از داده‌های آمارگیری پانلی خانواری انگلیس (British Household Panel Survey)، روش جانهی با الگوریتم EM با دو روش جانهی دیگر از نظر معیارهای مختلف مقایسه می‌شود. نتایج این مطالعه نشان می‌دهد که جانهی متغیر تحت بررسی در آمارگیری پانلی خانواری انگلیس با استفاده از الگوریتم EM وقتی که همبستگی بین دو دوره زیاد باشد، عملکرد بهتری دارد.

واژگان کلیدی: آمارگیری پانلی؛ بی‌پاسخی؛ کاهش پاسخگو؛ داده‌های ناتمام؛ داده‌های کامل؛ الگوریتم EM؛ جانهی با نمونه‌ی جدید؛ جانهی با میانگین مشاهده‌های مشابه.

دریافت: ۱۳۸۸/۱۱/۱۰، پذیرش: ۱۳۸۹/۶/۷

* نویسنده‌ی عهده‌دار مکاتبات

۱- مقدمه

هدف از آمارگیری، گرداوری اطلاعات مربوط به پارامترهای ناشناخته‌ی جامعه‌ی هدف می‌باشد. برای براورد این پارامترها در صورتی‌که مجموعه داده‌ها مستطیلی باشد و گم‌شدگی در آن وجود نداشته باشد، از روش‌های آماری استاندارد استفاده می‌شود. با توجه به این‌که یکی از خطاهای غیر نمونه‌گیری، خطای بی‌پاسخی است و معمولاً نمی‌توان از بروز این خطا جلوگیری کرد؛ بنا بر این نمی‌توان از روش‌های آماری استاندارد برای براورد پارامترهای جامعه استفاده کرد. در نتیجه برای تحلیل و استنباط نیاز به روش‌هایی برای تعدیل گم‌شدگی می‌باشد. وجود داده‌های گم‌شده یکی از مسئله‌های اساسی در آمارگیری‌های پانلی می‌باشد. این داده‌ها عمدتاً ناشی از کاهش پاسخگویان در دوره‌های دوم به بعد آمارگیری و یا عدم پاسخگویی به پاره‌ای از سئوال‌ها از سوی پاسخگویان می‌باشد. به‌ویژه هنگامی که داده‌های گم‌شده ناشی از کاهش نمونه در مطالعه باشد، مشکل‌های اساسی برای نتیجه‌گیری صحیح مطالعه فراهم می‌سازند، زیرا داده‌های گم‌شده می‌توانند قدرت استنباط آماری مطالعه را کاهش داده، باعث کاهش کارایی براوردگرها و ایجاد اریبی شوند. بنا بر این داده‌های گم‌شده به‌عنوان یک منبع بالقوه‌ی اریبی برای گرداوری داده‌ها در مطالعه‌های پانلی محسوب می‌شوند که می‌توانند منجر به نتایج گمراه‌کننده در تحلیل داده‌ها شوند. تعیین رهیافت مناسب برای رفتار با داده‌های گم‌شده و استفاده از استراتژی‌های مناسب هنگامی که میزان داده‌های گم‌شده زیاد است، بسیار حائز اهمیت می‌باشد.

روش‌های جانهی متفاوتی برای داده‌های گم‌شده در آمارگیری‌های پانلی استفاده می‌شود از جمله در پانل اقتصادی- اجتماعی آلمان از روش جانهی که توسط لیتل و سو [۸] گسترش یافته است، استفاده می‌کند. در آمارگیری نیروی انسانی و پویایی درامد کانادا از آخرین مقدار به دست آمده، به‌عنوان روش مقدماتی استفاده شده و در مورد داده‌های گم‌شده از سال قبل، روش نزدیک‌ترین همسایه به کار گرفته می‌شود. در آمارگیری پانلی پویایی درامد امریکا به‌طور کلی از روش بی‌درنگ برای جانهی داده‌های گم‌شده استفاده می‌شود [۴]. آمارگیری درامد و مشارکت برنامه‌ریزی امریکا، از دو روش جانهی استفاده می‌کند. در این آمارگیری، در بی‌پاسخی قلم اطلاعاتی از روش جانهی بی‌درنگ دنباله‌ای (Sequential Hot Deck Imputation) استفاده می‌شود [۱۰] و در

بی‌پاسخی دوره از روش جانهی طولی که روش منقول تصادفی (Random Carryover Method) نام نهاده شده است، استفاده می‌شود [۱۳].

در بخش دوم این مقاله، به معرفی آمارگیری پانلی پرداخته می‌شود. بخش سوم، ساختارهای گم‌شدگی به اختصار بیان شده و در بخش چهارم الگوریتم EM به‌عنوان روشی برای جانهی داده‌های گم‌شده معرفی می‌شود. بخش پنجم شامل طرح مطالعه و معیارهای مقایسه‌ی روش‌های جانهی استفاده شده می‌باشد، در آخر کاربردی از الگوریتم EM با استفاده از داده‌های پانلی خانواری انگلیس ارایه می‌شود.

۲- آمارگیری پانلی

برای براورد تغییر پارامترها در طی زمان نیاز به انتخاب روش مناسب آمارگیری است، لذا دو نوع روش آمارگیری در طی زمان، آمارگیری پانلی (طولی) و آمارگیری مقطعی وجود دارد. تفاوت این دو روش در این است که در آمارگیری مقطعی نمونه‌های مستقلی در مقاطع مختلف زمانی از جامعه‌ی هدف گزینش می‌شوند در حالی که در آمارگیری پانلی، نمونه‌ها ثابت و متشکل از واحدهایی است که در دوره‌ی اول در مطالعه حضور داشته‌اند.

یکی از مزیت‌های داده‌های پانلی این است که در آن داده‌های هر خوشه دارای همبستگی می‌باشند، این همبستگی معمولاً مثبت است. اگر براورد نرخ تغییر میانگین مد نظر باشد واریانس تغییر میانگین در آمارگیری‌های پانلی در دو دوره به‌صورت زیر محاسبه می‌شود:

$$\mathrm{var}(\bar{Y}_۲ - \bar{Y}_۱) = \mathrm{var}(\bar{Y}_۲) + \mathrm{var}(\bar{Y}_۱) - ۲\mathrm{cov}(\bar{Y}_۲, \bar{Y}_۱)$$

که در آن $\bar{Y}_t$ میانگین متغیر پاسخ در دوره‌ی tام ($t = ۱,۲$) آمارگیری است. مزیت دیگر آمارگیرهای طولی کاهش در هزینه‌های آمارگیری است. البته کار با داده‌های طولی پیچیده‌تر از داده‌های مقطعی است بنا بر این هزینه‌ی کار با آن‌ها زیادتر است. موزر و کالتون [۹] در بررسی‌های خود اظهار داشتند که انتخاب نمونه‌ی مناسب در آمارگیرهای پانلی نسبت به دیگر روش‌های آمارگیری دشوارتر است، زیرا اگر واحدهای نمونه‌ای مردم یک جامعه باشند تمایلی به این‌که رفتار آن‌ها در بازه‌های زمانی مورد بررسی قرار گیرد،

نداشته و نسبت به این امر واکنش نشان می‌دهند. این امر ممکن است سبب بی‌پاسخی در دوره‌های بعدی شود.

در آمارگیری پانلی نوعی بی‌پاسخی وجود دارد که به‌عنوان بی‌پاسخی دوره شناخته می‌شود به این صورت که از واحدهای نمونه‌ای در یک یا چند دوره‌ی آمارگیری و نه در تمام دوره‌ها هیچ‌گونه پاسخی دریافت نشود. کاهش پاسخگو حالت خاصی از بی‌پاسخی دوره است به این معنی که بعضی از واحدهای نمونه‌ای که در دور اول انتخاب می‌شوند از یک دوره تا انتهای مطالعه قابل دسترسی نباشند. لیتل و دیوید [۶]، سه نوع بی‌پاسخی دوره را معرفی کردند که شامل کاهش پاسخگو، ورود مجدد و ورود با تأخیر می‌باشند. بی‌پاسخی دوره معمولاً به‌عنوان مجموعه‌ای از بی‌پاسخی قلم اطلاعاتی قلمداد می‌شود لذا در این حالت جانهی مناسب است [۵].

۳ - ساختارهای گم‌شدگی

در این بخش به معرفی ساختارهای گم‌شدگی در آمارگیری‌های پانلی پرداخته می‌شود.

متغیر نشانگر پاسخ R_{it} به‌صورت زیر تعریف می‌کنیم ($t = ۱,۲$):

$$R_{it} = \begin{cases} ۱ & \text{اگر متغیر پاسخ برای فرد } i \text{ در دوره‌ی } t \text{ مشاهده شود} \\ ۰ & \text{اگر } Y_{it} \text{ گمشده باشد} \end{cases}$$

فرض کنید که $Y_{i۱}^{o}$ نمایانگر مقدار مشاهده‌شده‌ی واحد i ام در دور اول آمارگیری و $Y_{i۲}^{m}$ مقدار گم‌شده‌ی واحد i ام در دور دوم آمارگیری باشد. احتمال گم‌شدگی واحد i ام در دور دوم آمارگیری به‌شرط مقدار پاسخ مشاهده‌شده‌ی $Y_{i۱}^{o}$ و مقدار پاسخ گم‌شده‌ی $Y_{i۲}^{m}$ به‌صورت زیر نشان داده می‌شود:

$$P(R_{i۲} = ۰ | Y_{i۱}^{o}, Y_{i۲}^{m})$$

روبین [۱۱] ساختارهای گم‌شدگی را به‌صورت زیر بیان می‌کند:

وقتی که گم‌شدگی به مقدارهای گم‌شده و مشاهده‌شده وابسته نباشد، گم‌شدگی کاملاً تصادفی (Missing Completely at Random) است. در این حالت

$$P(R_{i۲}=۰|Y_{i۱}^{o},Y_{i۲}^{m})=P(R_{i۲}=۰)=p$$

به عبارت دیگر احتمال این‌که پاسخ واحد i ام در دور دوم که با نماد p نشان داده شده است، گم‌شده باشد به مقدار پاسخ مشاهده‌شده‌ی واحد i ام در دور اول یا مقدار پاسخ واحد i ام در دور دوم که گم‌شده است، بستگی ندارد.

در حالتی که گم‌شدگی متغیر پاسخ در دور دوم به‌دلیل پاسخی که در دور اول مشاهده‌شده رخ دهد، گم‌شدگی تصادفی (Missing at Random) است و

$$P(R_{i۲}=۰|Y_{i۱}^{o},Y_{i۲}^{m})=P(R_{i۲}=۰|Y_{i۱}^{o})=p_i^{o}$$

یعنی احتمال گم‌شدگی پاسخ واحد i ام در دور دوم وابسته به مقدار مشاهده‌شده‌ی واحد i ام در دور اول آمارگیری است که نماد p_i^{o} احتمال گم‌شدگی پاسخ واحد i ام در دور دوم که به داده‌های مشاهده‌شده وابسته است را نشان می‌دهد.

اگر احتمال مقادیر گم‌شده، به مقادیر مشاهده‌نشده‌ای که باید به دست آید، وابسته باشد گم‌شدگی را غیر تصادفی (Not Missing at Random) نامید، که در آن p_i^{m} احتمال گم‌شدگی واحد i ام که به داده‌های گم‌شده وابسته است را نشان می‌دهد.

$$P(R_{i۲}=۰|Y_{i۱}^{o},Y_{i۲}^{m})=p_i^{m}$$

۴- الگوریتم EM

الگوریتم EM [۳] یک روش معمول برای ماکسیمم کردن درستنمایی‌های پیچیده در برخورد با مسئله‌ی داده‌های ناتمام است. دامنه‌ی وسیعی از مسئله‌های آماری می‌تواند توسط این الگوریتم حل شود [۷]. از جمله کاربردهای این الگوریتم تعدیل اثر بی‌پاسخی می‌باشد.

دو فرض زیر برای استفاده از این الگوریتم در داده‌های گم‌شده مورد نیاز می‌باشند:

۱. پارامتر θ از پارامترهای فرایند داده‌های گم‌شده مستقل باشد، و

۲. داده‌های گم‌شده دارای ساختار گم‌شدگی تصادفی باشند.

این الگوریتم بر اساس ایده‌ای بی‌قاعده برای برخورد با داده‌های ناتمام فرمول‌بندی شده است، به‌طوری‌که:

۱. مقدارهای گم‌شده را توسط مقدارهای براورد شده جایگذاری می‌کند، و

۲. پارامترها را براورد می‌کند، و
۳. مقدارهای گم‌شده را با براوردهای جدید پارامترها دوباره براورد می‌کند، و
۴. مجدداً پارامترها را براورد می‌کند و تکرار این گام‌ها تا همگرایی ادامه می‌یابد.

براورد داده‌های گم‌شده در مرحله‌ی آخر به‌عنوان جانهی مقادیر گم‌شده استفاده می‌شود. برای استفاده از الگوریتم EM مجموعه داده‌ها باید به دو مجموعه داده تقسیم شود: اول مجموعه داده‌های مشاهده‌شده (داده‌های ناتمام) و دوم مجموعه داده‌های غیر قابل مشاهده (داده‌های کامل). در واقع مجموعه داده‌های ناتمام، مستقیماً مشاهده می‌شوند و مجموعه داده‌های کامل همراه با تعدادی داده‌ی گم‌شده می‌باشد. به بیان دیگر اگر y نماینده‌ی داده‌های ناتمام و x نماینده‌ی داده‌های کامل باشد، داریم $y = h(x)$ ، یعنی y کاملاً توسط x تعیین می‌شود، اما برعکس آن درست نیست.

فرض کنید که تابع‌های درستنمایی داده‌های ناتمام و داده‌های کامل به‌ترتیب با $L(\theta; y)$ و $L(\theta; x)$ نشان داده شوند که θ پارامتر توزیع تحت بررسی است. برای معرفی الگوریتم EM ابتدا رابطه‌ی زیر تعریف می‌شود:

$$Q(\theta'|\theta) = E[\log f(X|\theta')|y, \theta]$$

که θ پارامتر در مرحله‌ی قبل و θ' پارامتر در مرحله‌ی جدید را نشان می‌دهد. فرض می‌شود که برای تمام زوج‌های (θ', θ) تابع Q وجود دارد. همچنین فرض می‌شود که مقدار $f(x|\theta) > 0$ برای هر $\theta \in \Theta$ و x برقرار باشد. در این صورت گام‌های الگوریتم EM در مرحله‌ی $k+1$ ام به‌صورت زیر تعریف می‌شود:

گام E: مقدار $Q(\theta|\theta^{(k)})$ محاسبه می‌شود.

گام M: مقدار $\theta^{(k+1)}$ که $\theta \in \Theta$ طوری تعیین می‌شود که تابع $Q(\theta|\theta^{(k)})$ را ماکسیمم کند.

نکته‌ی قابل توجه این است که ماکسیمم‌سازی تحت اولین شناسه‌ی تابع Q صورت می‌گیرد. در اینجا ایده‌ی اصلی به این صورت است که برای تعیین کردن براورد ماکسیمم درستنمایی پارامتر θ نیاز به ماکسیمم کردن $\log f(x|\theta)$ می‌باشد؛ اما چون $\log f(x|\theta)$ به‌طور کامل مشخص نیست لذا به جای آن، امید ریاضی $\log f(x|\theta)$

به‌شرط داده‌های ناتمام y و $\theta^{(k)}$ ماکسیمم می‌شود. در واقع در گام E، این الگوریتم امید ریاضی شرطی داده‌های گم‌شده به شرط داده‌های مشاهده‌شده و پارامترهای براورد شده‌ی جاری را محاسبه می‌کند و سپس این امید ریاضی را برای داده‌های گم‌شده جایگزین می‌کند، که همان جانهی داده‌های گم‌شده می‌باشد. گام M پس از این‌که مقدارهای گم‌شده جایگزین شدند، مانند به دست آوردن براورد ML به‌سادگی محاسبه می‌شود.

۵- یک کاربرد

۵-۱- طرح مطالعه

آمارگیری پانلی خانواری انگلیس توسط مرکز مطالعات طولی انگلیس (The UK Longitudinal Studies Center) همراه با مؤسسه‌ی تحقیقات اقتصادی و اجتماعی دانشگاه اسکس (Institute for Social and Economic Research at the University of Essex) هر سال اجرا می‌شود [۱۲]. هدف اصلی این آمارگیری یافتن تغییرات اقتصادی و اجتماعی در سطح فرد و خانوار در انگلیس است. انتخاب مدل، پیشگویی و علت‌های این تغییرات در ارتباط با رده‌ای از متغیرهای اقتصادی-اجتماعی از جمله هدف‌های این آمارگیری است. این آمارگیری از سال ۱۹۹۱ تا کنون هر سال از ابتدای سپتامبر تا انتهای آوریل سال بعد اجرا می‌شود.

متغیرهای زیادی در آمارگیری پانلی خانواری انگلیس بررسی می‌شوند که در این مطالعه تنها از یک متغیر پیوسته (پرداخت ناخالص) در دوره‌ی هفدهم استفاده شده است. داده‌های این آمارگیری تا دوره‌ی هفدهم در دسترس می‌باشد. ما در این مطالعه تنها داده‌های دوره‌ی هفدهم را در نظر گرفتیم و چون بررسی اثر میزان همبستگی بین دو دوره‌ی متوالی مد نظر است لذا داده‌های دوره‌ی بعد را با در نظر گرفتن همبستگی‌های ($\rho = ۰/۳، ۰/۵، ۰/۷$) به‌طور مصنوعی ایجاد کردیم.

با توجه به این‌که داده‌های دوره‌ی هجدهم به‌طور مصنوعی با همبستگی‌های متفاوت با دوره‌ی قبل (دوره‌ی هفدهم) با استفاده از مدل $X_۲ = \rho X_۱ + Z$ ، که X_i نمایانگر متغیر مورد بررسی در دوره‌ی i ام ($i = ۱,۲$) است، تولید می‌شود و $Z \sim N\,(۰, (۱-\rho^۲)\mathrm{var}(X_۱))$ ، لذا توزیع متغیر مورد مطالعه در دوره‌ی هجدهم نیز نرمال می‌باشد.

جامعه‌ی هدف این مطالعه داده‌های دوره‌ی هفدهم که به اندازه‌ی ۶۷۶۰ است، فرض شد و از این جامعه نمونه‌هایی به‌ترتیب به اندازه‌های ۵۰، ۲۰۰ و ۵۰۰ به‌صورت تصادفی ساده بدون جایگذاری انتخاب شد. نرخ بی‌پاسخی برای این نمونه‌ها در دور دوم آمارگیری به ترتیب ۰/۰۵ و ۰/۱ در نظر گرفته شد. برای براورد توزیع نمونه‌گیری آماره‌ی تغییر میانگین در دو دور به اندازه‌ی ۱۰۰۰ مرتبه نمونه‌گیری مستقل از هر دوره صورت گرفت و براورد تغییر میانگین متغیر پاسخ و واریانس آن به دست آمد. فرض می‌شود که تمام واحدهای نمونه‌ای در دور اول مطالعه حضور دارند و کاهش پاسخگو در دور دوم به‌صورت تصادفی رخ می‌دهد.

سه روش جانهی وقتی که گم‌شدگی تصادفی در دور دوم مطالعه رخ دهد، بررسی می‌شود. این روش‌ها شامل جانهی با الگوریتم EM، جانهی با میانگین مشاهده‌های مشابه و جانهی با نمونه‌ی جدید [۱] می‌باشند. اکنون به معرفی جانهی داده‌های گم‌شده با استفاده از الگوریتم EM می‌پردازیم.

اگر متغیر مورد بررسی در دوره‌های اول و دوم آمارگیری را به‌صورت بردار تصادفی $\boldsymbol{X} = (X_1, X_2)^T$ نشان دهیم و فرض شود $\boldsymbol{X} \sim N_2(\mu, \Sigma)$ ، که $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ و

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

بردار داده‌های ناتمام $\boldsymbol{y}$ ، شامل تمام داده‌های دور اول و داده‌های مشاهده‌شده‌ی دور دوم آمارگیری است، که آن را به‌صورت $\boldsymbol{y} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_m^T, \boldsymbol{v}^T)^T$ که $\boldsymbol{v} = (x_{1,m+1}, \ldots, x_{1,m+m_1})^T$ نشان می‌دهیم. مقادیر گم‌شده نیز به‌صورت بردار $\boldsymbol{z} = (x_{2,m+1}, \ldots, x_{2,m+m_1})^T$ می‌باشند.

با توجه به این‌که توزیع جامعه نرمال دو متغیره است، گام‌های الگوریتم EM به‌صورت زیر می‌باشد:

گام E: تکرار $(k+1)$ ام در این گام بر اساس امید ریاضی شرطی لگاریتم درستنمایی داده‌های کامل به‌صورت

$$Q\left(\theta\middle|\theta^{(k)}\right)=E_{\theta^{(k)}}\left\{\log L_c(\theta)\middle|\boldsymbol{y}\right\}$$

می‌باشد، که $\theta^{(k)}$ مقدار θ را بعد از k امین تکرار الگوریتم EM و $L_c(\theta)$ معادله‌ی درستنمایی داده‌های کامل را نشان می‌دهند. برای محاسبه‌ی تابع Q تنها نیاز به محاسبه‌ی $E_{\theta^{(k)}}\left(X_{2,j}\middle|x_{1,j}\right)$ و $E_{\theta^{(k)}}\left(X_{2,j}^2\middle|x_{1,j}\right)$ برای $j=m+1,\ldots,m+m_1$ است، که به‌صورت زیر حاصل می‌شوند:

$$\begin{aligned}
&E_{\theta^{(k)}}\left(X_{2,j}\middle|x_{1,j}\right)=x_{2,j}^{(k)}=\mu_2^{(k)}+\left(\frac{\sigma_{12}^{(k)}}{\sigma_{11}^{(k)}}\right)(x_{1,j}-\mu_1^{(k)})\\
&E_{\theta^{(k)}}\left(X_{2,j}^2\middle|x_{1,j}\right)=x_{2,j}^{(k)2}+\sigma_{22.1}^{(k)},\qquad j=m+1,\ldots,m+m_1
\end{aligned}\tag{1}$$

گام M: با جایگذاری رابطه‌ی (۱) به جای مقادیر گم‌شده در معادله‌ی درستنمایی داده‌های کامل، براورد ML پارامترها در تکرار $(k+1)$ ام به دست می‌آید. سپس این دو گام تا همگرایی ادامه می‌یابند.

در روش جانهی با میانگین مشاهده‌های مشابه، داده‌های گم‌شده با استفاده از میانگین مشاهده‌های دوره‌ی دوم مطالعه که مقدار دوره‌ی اول آن‌ها برابر مقدار دوره‌ی اول مشاهده‌ی گم‌شده است، براورد می‌شود و اگر هیچ مقدار مشابهی برای داده‌های دور اول وجود نداشته باشد، مقدار پاسخ در دور اول جایگزین مقدار گم‌شده در دور دوم آمارگیری می‌شود.

در روش جانهی با نمونه‌ی جدید، یک نمونه‌ی تصادفی به اندازه‌ی m_1 از $N-n$ واحد جامعه انتخاب و مشاهده‌های دور دوم آن‌ها جایگزین مقادیر گم‌شده در نمونه‌ی اولیه می‌شوند.

یک جانهی خوب باید ویژگی‌های کلیدی آماری را از داده‌های کامل بازسازی کند. در ادامه برخی معیارها برای ارزیابی روش‌های جانهی ارایه می‌شوند.

۲-۵- معیارهای مقایسه

۲-۵-۱- قدر مطلق اریبی نسبی (Absolute Relative Bias)

اگر $\bar{x}_d$ میانگین براورد میانگین‌های تغییر ۱۰۰۰ نمونه‌ای باشد که از جامعه گرفته شده و مکانیسم گم‌شدگی در داده‌های دور دوم اعمال شده باشد، قدرمطلق تفاضل آن با میانگین تغییر پارامتر در جامعه μ_d در دورهای اول و دوم آمارگیری برای روش‌های مورد نظر و حالت‌های مختلف (اندازه‌ی نمونه، نرخ بی‌پاسخی و میزان همبستگی) محاسبه می‌شود. سپس از تقسیم قدرمطلق اریبی بر میانگین واقعی تغییر، μ_d ، قدرمطلق اریبی نسبی به دست می‌آید.

$$\widehat{ARB} = \frac{\left|\widehat{Bias}(\bar{x}_d)\right|}{\mu_d}$$

که در آن $\bar{x}_d = \frac{1}{1000}\sum_{h=1}^{1000}\bar{x}_{hd}$ و

$$\left|\widehat{\text{Bias}}(\bar{x}_d)\right| = \left|\hat{E}(\bar{x}_d) - \mu_d\right| = \left|\frac{1}{1000}\sum_{h=1}^{1000}\bar{x}_{hd} - \mu_d\right|$$

و $\bar{x}_{hd}$ میانگین تغییر در نمونه‌ی h ام می‌باشد.

۲-۵-۲- کارایی نسبی (Relative Efficiency)

اگر $\widehat{\text{var}}(\bar{x}_d)$ براورد واریانس براوردگر میانگین تغییر باشد، براورد میانگین توان دوم خطا از رابطه‌ی زیر محاسبه می‌شود:

$$\widehat{\text{MSE}}(\bar{x}_d) = \widehat{\text{var}}(\bar{x}_d) + [\widehat{\text{Bias}}(\bar{x}_d)]^2$$

و براورد واریانس براورد میانگین تغییر بر اساس ۱۰۰۰ تکرار نمونه‌گیری مستقل از جامعه از رابطه‌ی زیر به دست می‌آید:

$$\widehat{\text{var}}(\bar{x}_d) = \frac{1}{1000-1}\sum_{h=1}^{1000}(\bar{x}_{hd} - \bar{x}_d)^2$$

برای مقایسه‌ی دو به دوی روش‌های جانهی از معیار براورد کارایی نسبی (Estimated Relative Efficiency) برای تمام حالت‌های مورد نظر از رابطه‌ی زیر استفاده می‌شود:

$$\mathrm{ERE} = \frac{\widehat{\mathrm{MSE}}_{(1)}(\bar{x}_d)}{\widehat{\mathrm{MSE}}_{(2)}(\bar{x}_d)}$$

این معیار با مقداری کم‌تر از یک دلالت بر کارا بودن روش جانهی (۱) نسبت به روش جانهی (۲) دارد.

۳-۲-۵- درستی پیشگویی (Predictive Accuracy)

یک روش جانهی خوب باید از مقدارهای واقعی تا حد ممکن حفاظت کند، یعنی مقدارهای پیشگویی برای جانهی باید به مقدار واقعی «بسیار نزدیک» باشند. اگر این ویژگی به خوبی برقرار باشد مقدارهای $\hat{x}_i$ (مقدارهای جانهی شده) به مقدارهای x_i^* (مقدارهای واقعی) نزدیک خواهند بود. برای سنجش این معیار همبستگی پی‌یرسون میان x_i^* و $\hat{x}_i$ برای $i = 1, 2, \ldots, n$ بهترین ابزار می‌باشد [۲].

$$r_{\hat{x},x^*} = \frac{\sum_{i=1}^{n}(\hat{x}_i - \hat{\bar{x}})(x_i^* - \bar{x}^*)}{\sqrt{\sum_{i=1}^{n}(\hat{x}_i - \hat{\bar{x}})^2 \sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2}} \qquad (2)$$

که $\bar{x}^*$ و $\hat{\bar{x}}$ به‌ترتیب میانگین نمونه‌ای مقدارهای واقعی و جانهی می‌باشند. در یک جانهی خوب مقدار $r_{\hat{x},x^*}$ باید به ۱+ نزدیک باشد.

۴-۲-۵- درستی براورد (Estimation Accuracy)

در یک فرایند جانهی مناسب، براورد گشتاورهای مرتبه پایین باید به گشتاورهای داده‌های واقعی نزدیک باشند از این رو حفاظت از گشتاورهای توزیع تجربی مقدارهای واقعی از اهمیت زیادی برخوردار است.

برای ارزیابی درستی براورد، براوردهای پارامتر واریانس برای هر دو مقدارهای جانهی و مقدارهای واقعی محاسبه می‌شوند [۲].

قدر مطلق تفاضل واریانس (گشتاور مرتبه‌ی دوم): $M_{۲} = \left| M(x^{*۲}) - M(\hat{x}^{۲}) \right|$

$$M(\hat{x}^{۲}) = \hat{\text{var}}(\hat{\bar{x}}) = \frac{۱}{۱۰۰۰}\sum_{h=۱}^{۱۰۰۰}(\hat{\bar{x}}_h - \hat{\bar{x}})^{۲}$$

$$M(x^{*۲}) = \hat{\text{var}}(\bar{x}^{*}) = \frac{۱}{۱۰۰۰}\sum_{h=۱}^{۱۰۰۰}(\bar{x}_h^{*} - \bar{x}^{*})^{۲}$$

که در آن $\hat{\bar{x}}$ و $\bar{x}^{*}$ به‌ترتیب میانگین کل مقدارهای جانهی و واقعی و $\hat{\bar{x}}_h$ و $\bar{x}_h^{*}$ میانگین مقدارهای جانهی و واقعی در تکرار h ام می‌باشند. فرایندی خوب است که قدر مطلق تقاضل واریانس آن در مقایسه با سایر روش‌های جانهی کم‌تر باشد.

۵-۲-۵- درستی پیشگویی تغییر (Predictive Accuracy of Change)

معیار درستی پیشگویی تغییر میزان حفاظت بین دوره‌های آمارگیری را مشخص می‌کند. در این حالت بزرگی و کوچکی همبستگی ملاک نمی‌باشد بلکه معیار اصلی نزدیک بودن مقدار همبستگی داده‌های واقعی در دو دوره به همبستگی داده‌های جانهی شده می‌باشد.

قدرمطلق تفاضل همبستگی‌ها به‌صورت زیر تعریف می‌شود:

$$\text{درستی پیشگویی تغییر بین دوره‌های اول و دوم آمارگیری} = \left| r_{x_{۱}^{*},\hat{x}_{۲}} - r_{x_{۱}^{*},x_{۲}^{*}} \right|$$

که در آن $r_{x_{۱}^{*},x_{۲}^{*}}$ براورد ضریب همبستگی مقدارهای واقعی دوره‌های اول و دوم آمارگیری و $r_{x_{۱}^{*},\hat{x}_{۲}}$ ضریب همبستگی داده‌های واقعی دور اول و داده‌های جانهی‌شده‌ی دور دوم آمارگیری می‌باشند که از رابطه‌ی (۲) محاسبه می‌شوند.

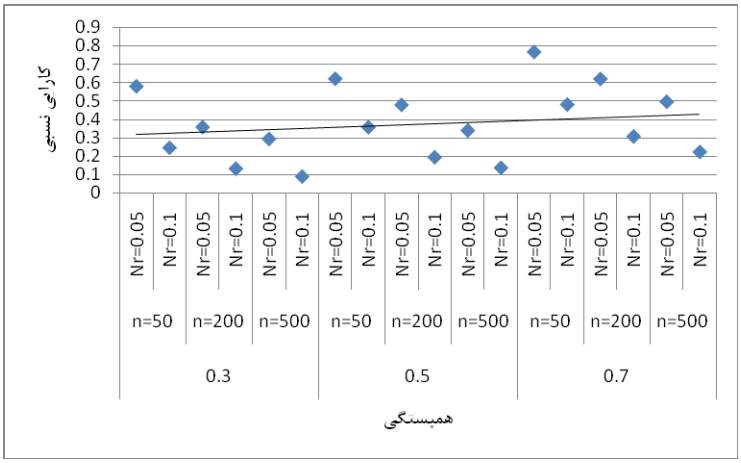
۵-۳- یافته‌های مطالعه

با توجه به آماره‌های جدول ۱ مشخص می‌شود که قدرمطلق اریبی نسبی براورد تغییر میانگین در دو دوره‌ی متوالی در روش جانهی با الگوریتم EM و نمونه‌ی جدید کم‌تر از روش جانهی با میانگین مشاهده‌های مشابه است. روش جانهی با نمونه‌ی جدید در همه‌ی حالت‌ها عمل‌کرد بهتری نسبت به دو روش دیگر دارد.

جدول ۱- قدر مطلق اریبی نسبی براورد تغییر میانگین دردو دوره‌ی متوالی در سه روش جانهی

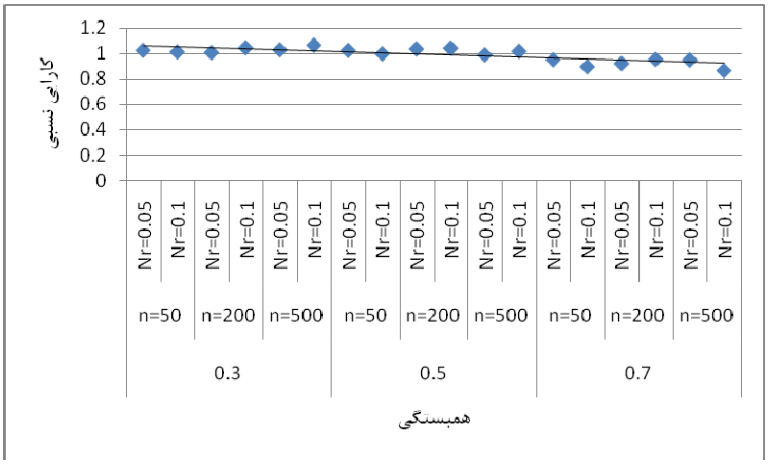
همبستگی	اندازه‌ی نمونه	نرخ بی‌پاسخی	الگورتیم EM	میانگین مشاهده‌های مشابه	نمونه‌ی جدید
۰/۳	۵۰	۰/۰۵	۰/۰۱۵۴۰۱۴۶	۰/۰۵۶۳۴۲۱۴	۰/۰۱۳۴۰۳۱۵
		۰/۱	۰/۰۱۵۸۶۶۳۸	۰/۱۰۱۸۵۶۱	۰/۰۱۳۵۲۰۶۴
	۲۰۰	۰/۰۵	۰/۰۱۷۲۰۲۳۴	۰/۰۴۷۵۶۰۳۶	۰/۰۱۳۵۸۹۲۴
		۰/۱	۰/۰۱۸۲۰۴۰۴	۰/۰۸۰۴۳۸۶۵	۰/۰۱۲۶۲۵۸۸
	۵۰۰	۰/۰۵	۰/۰۱۶۲۸۵۹	۰/۰۳۹۶۲۳۳۷	۰/۰۱۴۱۱۰۶۵
		۰/۱	۰/۰۱۶۱۲۸۰۶	۰/۰۶۵۹۴۵۰۵	۰/۰۱۳۰۱۶۵۱
۰/۵	۵۰	۰/۰۵	۰/۰۰۳۴۱۲۳۲۷	۰/۰۴۴۱۳۲۵۳	۰/۰۰۰۸۱۰۱۲۲۵
		۰/۱	۰/۰۰۳۱۳۳۹۲۱	۰/۰۹۰۱۳۰۰۴	۰/۰۰۲۹۲۳۱۸۵
	۲۰۰	۰/۰۵	۰/۰۰۵۶۹۸۷۷۷	۰/۰۳۴۸۷۶۲۹	۰/۰۰۱۷۰۲۷
		۰/۱	۰/۰۰۷۶۶۳۱۹۴	۰/۰۶۹۱۵۴۰۴	۰/۰۰۱۴۷۳۷۰۷
	۵۰۰	۰/۰۵	۰/۰۰۵۲۹۵۱	۰/۰۲۸۷۹۳۵۱	۰/۰۰۰۳۰۵۳
		۰/۱	۰/۰۰۳۹۷۳۴۴۳	۰/۰۵۳۷۰۶۰۷	۰/۰۰۱۸۷۷۲۳
۰/۷	۵۰	۰/۰۵	۰/۰۰۶۴۸۰۱۸۳	۰/۰۴۷۴۲۵۵۲	۰/۰۰۴۶۶۰۶۲
		۰/۱	۰/۰۰۴۹۱۸۶۱۴	۰/۰۹۱۴۴۷۳۷	۰/۰۰۲۱۲۵۵۰۲
	۲۰۰	۰/۰۵	۰/۰۰۵۷۲۹۶۲۹	۰/۰۳۳۸۴۷۷۵	۰/۰۰۳۴۰۴۹
		۰/۱	۰/۰۰۹۵۴۱۰۴۶	۰/۰۶۸۳۵۵۲۴	۰/۰۰۲۴۴۶۵۱۶
	۵۰۰	۰/۰۵	۰/۰۰۲۷۷۱۱۴	۰/۰۲۵۵۷۰۴۲	۰/۰۰۰۲۷۲۳۸۵۷
		۰/۱	۰/۰۰۳۷۸۴۷۴۸	۰/۰۵۲۷۲۴۲۲	۰/۰۰۱۷۳۵۴۴۳

با ملاحظه‌ی کارایی نسبی روش جانهی با الگوریتم EM نسبت به میانگین مشاهده‌های مشابه در اندازه‌های نمونه‌ی ۵۰ و ۲۰۰ و ۵۰۰ و تمامی حالت‌های کاهش پاسخگو مشخص می‌شود که جانهی با الگوریتم EM کاراتر از جانهی با میانگین مشاهده‌های مشابه است، همچنین از شکل ۱ مشخص است که با افزایش نمونه و افزایش نرخ بی‌پاسخی جانهی با الگوریتم EM نسبت به جانهی با میانگین مشاهده‌های مشابه کاراتر می‌شود.



شکل ۱- کارایی نسبی روش جانهی با الگوریتم EM نسبت به میانگین مشاهده‌های مشابه برای تمام حالت‌های مورد نظر

معیار کارایی نسبی در شکل ۲ نشان می‌دهد که دو روش جانهی با الگوریتم EM و جانهی با نمونه‌ی جدید وقتی که همبستگی بین دو دوره ۰/۳ و ۰/۵ باشد، تفاوت چندانی ندارند، در حالی که وقتی همبستگی بین دو دوره ۰/۷ باشد روش جانهی با الگوریتم EM بهتر از روش جانهی با نمونه‌ی جدید است.



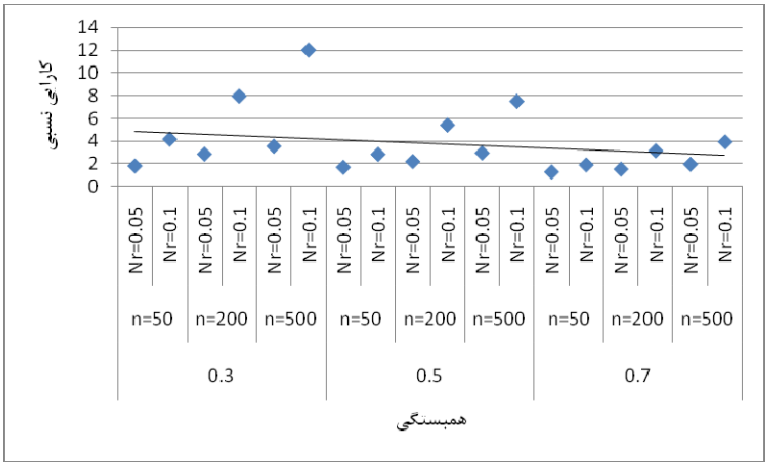
شکل ۲- کارایی نسبی روش جانهی با الگوریتم EM نسبت به نمونه‌ی جدید برای تمام حالت‌های مورد نظر

با مقایسه‌ی دو روش جانهی با میانگین مشاهده‌های مشابه و نمونه‌ی جدید از نظر معیار کارایی نسبی مشخص است که روش جانهی با نمونه‌ی جدید نسبت به میانگین مشاهده‌های مشابه کاراتر است، همچنین با افزایش همبستگی بین دو دوره میزان کارایی روش جانهی با نمونه‌ی جدید نسبت به میانگین مشاهده‌های مشابه کم‌تر می‌شود.

همان‌طور که در جدول ۲ نشان داده شده است مقدار همبستگی بین دوره‌ی دوم نمونه‌ی بدون بی‌پاسخی و دوره‌ی دوم نمونه‌ی جانهی‌شده در روش جانهی با الگوریتم EM نسبت به دو روش جانهی با نمونه‌ی جدید و میانگین مشاهده‌های مشابه بیش‌تر می‌باشد. بعد از روش الگوریتم EM، روش میانگین مشاهده‌های مشابه نتایج خوبی را نشان داده است در حالی که روش جانهی با نمونه‌ی جدید اصلاً عمل‌کرد خوبی نداشته است، همچنین با افزایش نرخ بی‌پاسخی مقدار همبستگی کاهش می‌یابد.

در مورد اثر نرخ بی‌پاسخی و همبستگی بر اساس معیار درستی پیشگویی روش جانهی با الگوریتم EM می‌توان گفت:

- با افزایش نرخ بی‌پاسخی درستی پیشگویی روش جانهی با الگوریتم EM کاهش می‌یابد.
- با افزایش همبستگی بین دو دوره میزان درستی پیشگویی روش جانهی با الگوریتم EM افزایش می‌یابد.



شکل ۳- کارایی نسبی روش جانهی با میانگین مشاهده‌های مشابه نسبت به نمونه‌ی جدید برای تمام حالت‌های مورد نظر

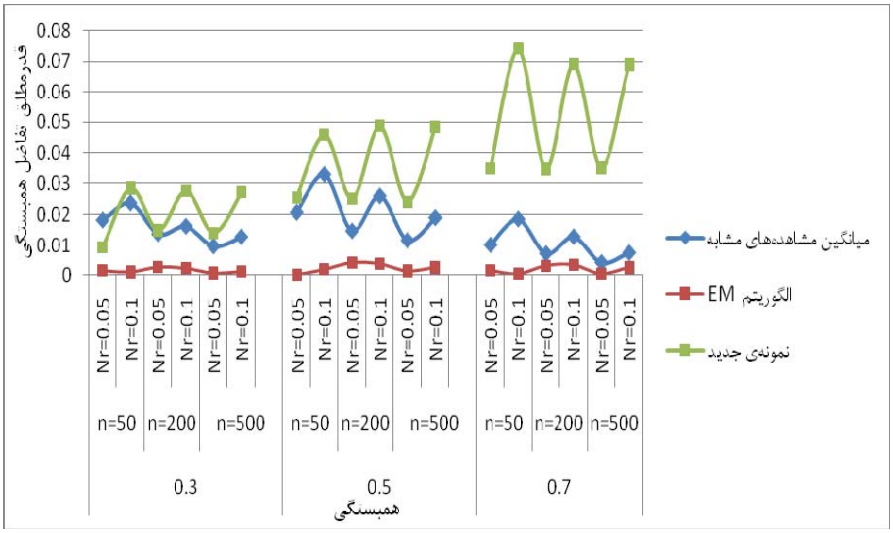
جدول ۲- مقایسه‌ی درستی پیشگویی سه روش جانهی

همبستگی	اندازه‌ی نمونه	نرخ بی‌پاسخی	الگورتیم EM	میانگین مشاهده‌های مشابه	نمونه‌ی جدید
۰/۳	۵۰	۰/۰۵	۰/۹۴۸۲۰۰۶	۰/۸۰۲۵۸۴۵	۰/۳۴۷۰۵۳
		۰/۱	۰/۹۰۷۲۴۸	۰/۶۷۴۵۵۲۹	۰/۱۷۲۷۹۱۶
	۲۰۰	۰/۰۵	۰/۹۴۴۴۱۷	۰/۸۱۷۴۰۷	۰/۰۸۸۷۷۸۳۴
		۰/۱	۰/۸۹۹۲۸۲۵	۰/۶۹۹۹۰۱	۰/۰۳۶۲۴۵۶۵
	۵۰۰	۰/۰۵	۰/۹۵۲۰۲۷۵	۰/۸۴۷۸۱۵۳	۰/۰۳۵۵۹۶۲۳
		۰/۱	۰/۹۰۲۴۱۵۸	۰/۷۳۴۲۸۰۵	۰/۰۱۶۶۸۷۶۸
۰/۵	۵۰	۰/۰۵	۰/۹۶۰۵۴۳	۰/۸۹۰۴۱۷	۰/۳۴۲۰۳۴۹
		۰/۱	۰/۹۱۹۳۶۶	۰/۷۹۹۹۲۹۱	۰/۱۶۲۱۷۷۱
	۲۰۰	۰/۰۵	۰/۹۵۵۲۵۴	۰/۹۰۳۷۸۲۲	۰/۰۹۱۰۲۷۰۲
		۰/۱	۰/۹۱۷۲۵۸۷	۰/۸۲۴۴۸۶	۰/۰۴۳۱۱۲۲۸
	۵۰۰	۰/۰۵	۰/۹۶۰۰۷۱۳	۰/۹۱۶۸۶۸۲	۰/۰۳۵۹۶۳۵۸
		۰/۱	۰/۹۱۸۵۷۳۵	۰/۸۴۶۲۲۴۵	۰/۰۱۶۴۱۵۳۱
۰/۷	۵۰	۰/۰۵	۰/۹۷۲۳۳۵۷	۰/۹۴۹۳۲۸۵	۰/۳۳۶۳۱۴۲
		۰/۱	۰/۹۴۴۸۰۵۷	۰/۹۰۴۳۰۴۶	۰/۱۶۳۷۸۹۲
	۲۰۰	۰/۰۵	۰/۹۶۹۱۲۵۵	۰/۹۵۶۰۳۱۴	۰/۰۸۳۴۸۰۱۳
		۰/۱	۰/۹۴۴۸۱۵۳	۰/۹۱۶۸۸۵۶	۰/۰۴۱۱۳۹۶۶
	۵۰۰	۰/۰۵	۰/۹۷۳۷۲۳۳	۰/۹۶۲۰۸۵۴	۰/۰۳۴۷۲۹۳۳
		۰/۱	۰/۹۴۵۴۱۵۷	۰/۹۲۶۰۰۲۳	۰/۰۱۶۸۳۱۶۳

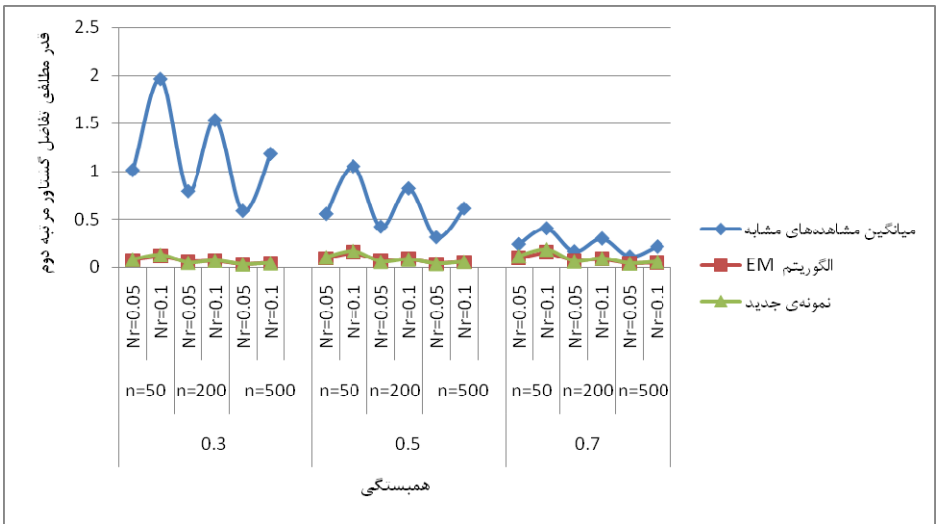
معیار درستی پیشگویی تغییر (شکل ۴) در روش جانهی با الگوریتم EM نسبت به دو روش جانهی با نمونه‌ی جدید و میانگین مشاهده‌های مشابه بهتر می‌باشد. شایان ذکر است که این معیار در روش جانهی با نمونه‌ی جدید عمل‌کرد خوبی ندارد. در این روش قدر مطلق تفاضل همبستگی‌ها با افزایش همبستگی بین دو دوره بیش‌تر می‌شود.

در شکل ۵ با مقایسه‌ی معیار درستی براورد در سه روش جانهی مشخص می‌شود که این معیار در روش جانهی با الگوریتم EM و جانهی با نمونه‌ی جدید نسبت به روش

جانهی با میانگین مشاهده‌های مشابه کم‌تر است، همچنین با افزایش نرخ بی‌پاسخی میزان درستی براورد کم‌تر می‌شود.



شکل ۴- مقایسه‌ی درستی پیشگویی تغییر سه روش جانهی



شکل ۵- مقایسه‌ی درستی براورد سه روش جانهی

لذا روش جانهی با الگوریتم EM از نظر معیار درستی براورد در آمارگیری پانلی خانواری انگلیس نسبت به دو روش جانهی با میانگین مشاهده‌های مشابه و نمونه‌ی جدید عمل‌کرد بهتری دارد. میزان درستی براورد با کاهش نرخ بی‌پاسخی و افزایش اندازه‌ی نمونه بیش‌تر می‌شود.

۶- خلاصه

در این مقاله ابتدا آمارگیری پانلی، انواع گم‌شدگی و ساختارهای گم‌شدگی در آمارگیری پانلی بیان شد سپس الگوریتم EM به‌عنوان روشی برای جانهی داده‌های گمشده در آمارگیری پانلی دو دوره‌ای وقتی که توزیع متغیر مورد نظر نرمال دومتغیره باشد، ارایه شد و روش جانهی با الگوریتم EM با دو روش جانهی میانگین مشاهده‌های مشابه و نمونه‌ی جدید مقایسه شد. با مطالعه‌ی شبیه‌سازی انجام‌شده، نتایج زیر حاصل می‌شود:

بر اساس معیارهای مقایسه‌ی قدر مطلق اریبی نسبی براورد میانگین تغییر و همچنین کارایی آن، وقتی که همبستگی بین دو دوره زیاد باشد روش جانهی با الگوریتم EM نسبت به دو روش جانهی با نمونه‌ی جدید و میانگین مشاهده‌های مشابه بهتر عمل می‌کند. همچنین طبق معیارهای درستی پیشگویی، درستی براورد و درستی پیشگویی تغییر روش جانهی با الگوریتم EM نیز نسبت به دو روش دیگر عمل‌کرد بهتری دارد (وقتی که براورد میانگین تغییر مد نظر نباشد).

در پایان با توجه به این‌که همبستگی متغیر پرداخت ناخالص بین دو دوره‌ی شانزدهم و هفدهم در داده‌های واقعی زیاد می‌باشد، لذا بر اساس معیارهای محاسبه‌شده و نتایج موجود، به نظر می‌رسد که روش جانهی با الگوریتم EM برای جانهی داده‌های گم‌شده‌ی متغیر پرداخت ناخالص در آمارگیری پانلی خانواری انگلیس، مفید باشد.

مرجع‌ها

[۱] نواب‌پور، حمیدرضا (در حال چاپ). مقایسه‌ی دو روش جانهی برای براورد نرخ تغییر میانگین در مطالعه‌های طولی با نمونه‌ی کاهیده. *مجله‌ی بررسی‌های آمار رسمی ایران.*

[2] Chambers, R. (2000). Evaluation Criteria for Statistical Editing and Imputation. *Working Paper for the Erudite Project on the Development and*

Evaluation of New Methods for Editing and Imputation, University of Southampton, Southampton, UK.

[3] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society, Series B, Methodological*, **39**, 1-38.

[4] Hofferth, S., Stafford, F.P., Yeung, W.J., Duncan, G.J., Hill, M.S., Lepkowski, J. abd Morgan, J.N. (1998). A Panel Study of Income Dynamics: Procedures and Codebooks. Guide to the 1993 Interviewing Year, *Institute for Social Research,* The University of Michigan.

[5] Kalton, G. (1986). Handling Wave Nonresponse in Panel Surveys, *Journal of Official Statistics*, **2**, 303-314.

[6] Little, R.J.A. and David, M.H. (1983). Weighting Adjustments for Non-response in Panel Surveys, *Working Paper,* U.S. Bureau of the Census, Washington D.C.

[7] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

[8] Little, R.J.A. and Su, H.L. (1989). *Item Non-Response in Panel Surveys.* Wiley, New York.

[9] Moser, C.A. and Kalton, G. (1979). *Survey Methods in Social Investigation*. Second Edition, London, Heinemann Educational Books.

[10] Pennell, S.G. (1993). Cross-Sectional Imputation and Longitudinal Editing Procedures in the Survey of Income and Program Participation, *Institute for Social Research,* The University of Michigan.

[11] Rubin, D.B. (1976). Inference and missing data, *Biometrika,* **63**, 581-592.

[12] Taylor, M.F., Brice, J., Buck, N. and Prentice-Lane, E. (2009). British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices. *Colchester: University of Essex*.

[13] Williams, T.R. and Bailey, L. (1996). Compensating for missing wave data in the survey of income and program participation (SIPP), proceedings of the survey research methods section, *American Statistical Association*, 305-310.

آسیه رشیدی‌نژاد
کارشناس ارشد آمار
تهران، خیابان شهید بهشتی، نبش احمد قصیر، دانشکده‌ی اقتصاد دانشگاه علامه‌طباطبایی، گروه آمار.
رایانشانی: *assiehrashidi@yahoo.com*

حمیدرضا نواب‌پور
دکتری آمار
تهران، خیابان شهید بهشتی، نبش احمد قصیر، دانشکده‌ی اقتصاد دانشگاه علامه‌طباطبایی، گروه آمار.
رایانشانی: *h.navvabpour@srtc.ac.ir*