

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

آسیه رشیدی‌نژاد* و حمیدرضا نواب‌پور

دانشگاه علامه طباطبائی

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 برآورد 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 برآوردها 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، گردآوری اطلاعات مربوط به پارامترهای ناشناخته‌ی جامعه‌ی هدف می‌باشد. برای برآورد 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 برآورد 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 هنگامی‌که 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 برآوردها و ایجاد اربیبی شوند. بنا بر این داده‌های گم‌شده به‌عنوان یک منبع بالقوه‌ی اربیبی برای گردآوری 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 هنگامی‌که 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 درآمد 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 درآمد امریکا به‌طور کلی از روش بی‌درنگ برای جانهی داده‌های گم‌شده استفاده می‌شود [۴]. آمارگیری درآمد 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 ارائه می‌شود.

۲- آمارگیری پانلی

برای برآورد 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 برآورد نرخ تغییر میانگین مد نظر باشد واریانس تغییر میانگین در آمارگیری‌های پانلی در دو دوره به‌صورت زیر محاسبه می‌شود:

$$\text{var}(\bar{Y}_2 - \bar{Y}_1) = \text{var}(\bar{Y}_2) + \text{var}(\bar{Y}_1) - 2\text{cov}(\bar{Y}_2, \bar{Y}_1)$$

که در آن \bar{Y}_t میانگین متغیر پاسخ در دوره‌ی t ام ($t = 1, 2$) آمارگیری است. مزیت دیگر آمارگیری‌های طولی کاهش در هزینه‌های آمارگیری است. البته کار با داده‌های طولی پیچیده‌تر از داده‌های مقطعی است بنا بر این هزینه‌ی کار با آن‌ها زیاده‌تر است. موزر و کالتون [۹] در بررسی‌های خود اظهار داشتند که انتخاب نمونه‌ی مناسب در آمارگیری‌های 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 ($t = 1, 2$):

$$R_{it} = \begin{cases} 1 & \text{اگر متغیر پاسخ برای فرد } i \text{ در دوره } t \text{ مشاهده شود} \\ 0 & \text{اگر } Y_{it} \text{ گم‌شده باشد} \end{cases}$$

فرض کنید که Y_{i1}^o نمایانگر مقدار مشاهده‌شده‌ی واحد i ام در دور اول آمارگیری و Y_{i2}^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_{i1}^o و مقدار پاسخ گم‌شده‌ی Y_{i2}^m به‌صورت زیر نشان داده می‌شود:

$$P(R_{i2} = 0 | Y_{i1}^o, Y_{i2}^m)$$

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

$$P(R_{i_2} = \circ | Y_{i_1}^o, Y_{i_2}^m) = P(R_{i_2} = \circ) = p$$

به عبارت دیگر احتمال این که پاسخ واحد i ام در دور دوم که با نماد p نشان داده شده است، گم شده باشد به مقدار پاسخ مشاهده شده‌ی واحد i ام در دور اول یا مقدار پاسخ واحد i ام در دور دوم که گم شده است، بستگی ندارد.

در حالتی که گم شدگی متغیر پاسخ در دور دوم به دلیل پاسخی که در دور اول مشاهده شده رخ دهد، گم شدگی تصادفی (Missing at Random) است و

$$P(R_{i_2} = \circ | Y_{i_1}^o, Y_{i_2}^m) = P(R_{i_2} = \circ | Y_{i_1}^o) = p_i^o$$

یعنی احتمال گم شدگی پاسخ واحد i ام در دور دوم وابسته به مقدار مشاهده شده‌ی واحد i ام در دور اول آمارگیری است که نماد p_i^o احتمال گم شدگی پاسخ واحد i ام در دور دوم که به داده‌های مشاهده شده وابسته است را نشان می‌دهد.

اگر احتمال مقادیر گم شده، به مقادیر مشاهده نشده‌ای که باید به دست آید، وابسته باشد گم شدگی را غیر تصادفی (Not Missing at Random) نامید، که در آن p_i^m احتمال گم شدگی واحد i ام که به داده‌های گم شده وابسته است را نشان می‌دهد.

$$P(R_{i_2} = \circ | Y_{i_1}^o, Y_{i_2}^m) = p_i^m$$

۴- الگوریتم EM

الگوریتم EM [۳] یک روش معمول برای ماکسیم 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 گم شده مورد نیاز می‌باشند:

۱. پارامتر θ از پارامترهای فرایند داده‌های گم شده مستقل باشد، و
 ۲. داده‌های گم شده دارای ساختار گم شدگی تصادفی باشند.
- این الگوریتم بر اساس ایده‌ای بی‌قاعده برای برخورد با داده‌های ناتمام فرمول‌بندی شده است، به طوری که:

۱. مقدارهای گم شده را توسط مقدارهای برآورد شده جایگذاری می‌کند، و

۲. پارامترها را برآورد می‌کند، و
۳. مقدارهای گم‌شده را با برآوردهای جدید پارامترها دوباره برآورد می‌کند، و
۴. مجدداً پارامترها را برآورد می‌کند و تکرار این گام‌ها تا همگرایی ادامه می‌یابد.
- برآورد 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 $L(\theta; x)$ و $L(\theta; y)$ 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 برآورد 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 $\log f(x|\theta)$

به شرط 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 برآورد 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 این که مقادیرهای گم‌شده جایگزین شدند، مانند به دست آوردن برآورد 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 $(\rho = 0/3, 0/5, 0/7)$ به‌طور مصنوعی ایجاد کردیم.

با توجه به این که داده‌های دوره‌ی هجدهم به‌طور مصنوعی با همبستگی‌های متفاوت با دوره‌ی قبل (دوره‌ی هفدهم) با استفاده از مدل $X_t = \rho X_{t-1} + Z_t$ ، که X_i نمایانگر متغیر مورد بررسی در دوره‌ی i ام ($i = 1, 2$) است، تولید می‌شود و $Z \sim N(0, (1 - \rho^2) \text{var}(X_{t-1}))$ ، لذا توزیع متغیر مورد مطالعه در دوره‌ی هجدهم نیز نرمال می‌باشد.

جامعه‌ی هدف این مطالعه داده‌های دوره‌ی هفدهم که به اندازه‌ی 676° است، فرض شد و از این جامعه نمونه‌هایی به ترتیب به اندازه‌های 5° ، 20° و 50° به صورت تصادفی ساده بدون جایگذاری انتخاب شد. نرخ بی‌پاسخی برای این نمونه‌ها در دور دوم آمارگیری به ترتیب $5/0^\circ$ و $1/0^\circ$ در نظر گرفته شد. برای برآورد توزیع نمونه‌گیری آماره‌ی تغییر میانگین در دو دور به اندازه‌ی 1000 مرتبه نمونه‌گیری مستقل از هر دوره صورت گرفت و برآورد 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 به صورت 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 به صورت بردار تصادفی $X = (X_1, X_2)^T$ نشان دهیم و فرض شود $X \sim N_2(\mu, \Sigma)$ ، که $\mu = (\mu_1, \mu_2)^T$ و

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

بردار داده‌های ناتمام y ، شامل تمام داده‌های دور اول و داده‌های مشاهده‌شده‌ی دور دوم آمارگیری است، که آن را به صورت $y = (x_1^T, \dots, x_m^T, v^T)^T$ که

$v = (x_{1,m+1}, \dots, x_{1,m+m})^T$ نشان می‌دهیم. مقادیر گم‌شده نیز به صورت بردار

$$z = (x_{2,m+1}, \dots, x_{2,m+m})^T$$

می‌باشند.

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

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

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

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 $E_{\theta^{(k)}}\left(X_{r,j} \mid x_{1,j}\right)$ و $E_{\theta^{(k)}}\left(X_{r,j}^2 \mid x_{1,j}\right)$ برای $j = m+1, \dots, m+m_1$ است، که به صورت زیر حاصل می‌شوند:

$$(1) \quad E_{\theta^{(k)}}\left(X_{r,j} \mid x_{1,j}\right) = x_{r,j}^{(k)} = \mu_r^{(k)} + \left(\frac{\sigma_{r2}^{(k)}}{\sigma_{11}^{(k)}}\right)(x_{1,j} - \mu_1^{(k)})$$

$$E_{\theta^{(k)}}\left(X_{r,j}^2 \mid x_{1,j}\right) = x_{r,j}^{(k)2} + \sigma_{rr,1}^{(k)}, \quad j = m+1, \dots, m+m_1$$

گام M: با جایگذاری رابطه‌ی (۱) به جای مقادیر گم‌شده در معادله‌ی درستنمایی داده‌های کامل، برآورد 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، برآورد 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 ارائه می‌شوند.

۵-۲- معیارهای مقایسه

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

اگر \bar{x}_d میانگین برآورد 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 دوره‌های 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.

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

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

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

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

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

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

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

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

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

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

$$ERE = \frac{\widehat{MSE}_{(s)}(\bar{x}_d)}{\widehat{MSE}_{(r)}(\bar{x}_d)}$$

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 $i = 1, 2, \dots, n$ بهترین ابزار می‌باشد [۲].

$$(۲) \quad r_{\hat{x}, x^*} = \frac{\sum_{i=1}^n (\hat{x}_i - \hat{\bar{x}})(x_i^* - \bar{x}^*)}{\sqrt{\sum_{i=1}^n (\hat{x}_i - \hat{\bar{x}})^2 \sum_{i=1}^n (x_i^* - \bar{x}^*)^2}}$$

که $\hat{\bar{x}}$ و \bar{x}^* به ترتیب 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 برآورد (Estimation Accuracy)

در یک فرایند جانهی مناسب، برآورد 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 برآورد، برآوردهای پارامتر واریانس برای هر دو مقدارهای جانهی و مقدارهای واقعی محاسبه می‌شوند [۲].

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

$$M(\hat{x}^{\tau}) = \text{var}(\hat{x}) = \frac{1}{1000} \sum_{h=1}^{1000} (\hat{x}_h - \hat{\bar{x}})^{\tau}$$

$$M(x^{*\tau}) = \text{var}(\bar{x}^*) = \frac{1}{1000} \sum_{h=1}^{1000} (\bar{x}_h^* - \bar{x}^*)^{\tau}$$

که در آن \hat{x} و \bar{x}^* به ترتیب میانگین کل مقدارهای جانهی و واقعی و \hat{x}_h و \bar{x}_h^* میانگین مقدارهای جانهی و واقعی در تکرار h ام می‌باشند. فرایندی خوب است که قدر مطلق تفاضل 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. قدر مطلق تفاضل همبستگی‌ها به صورت زیر تعریف می‌شود:

$$\left| r_{x_1^*, \hat{x}_2^*} - r_{x_1^*, x_2^*} \right| = \text{درستی پیشگویی تغییر بین دوره‌های اول و دوم آمارگیری}$$

که در آن $r_{x_1^*, x_2^*}$ برآورد ضریب همبستگی مقدارهای واقعی دوره‌های اول و دوم آمارگیری و $r_{x_1^*, \hat{x}_2^*}$ ضریب همبستگی داده‌های واقعی دور اول و داده‌های جانهی شده‌ی دور دوم آمارگیری می‌باشند که از رابطه‌ی (۲) محاسبه می‌شوند.

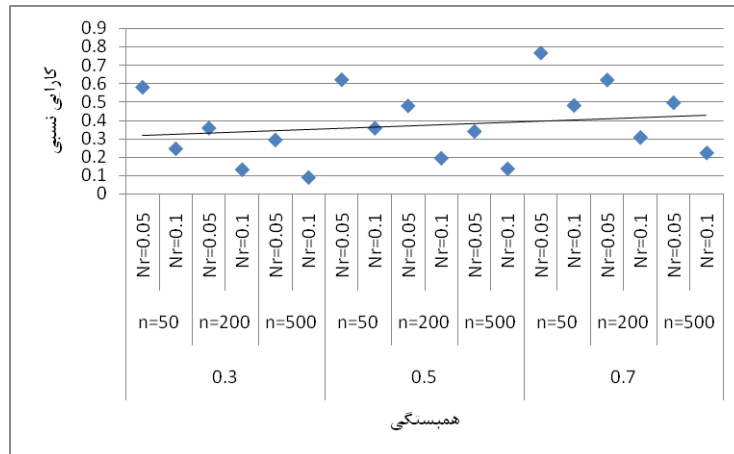
۵-۳- یافته‌های مطالعه

با توجه به آماره‌های جدول ۱ مشخص می‌شود که قدر مطلق اریبی نسبی برآورد 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 برآورد تغییر میانگین در دو دوره‌ی متوالی در سه روش جانهی

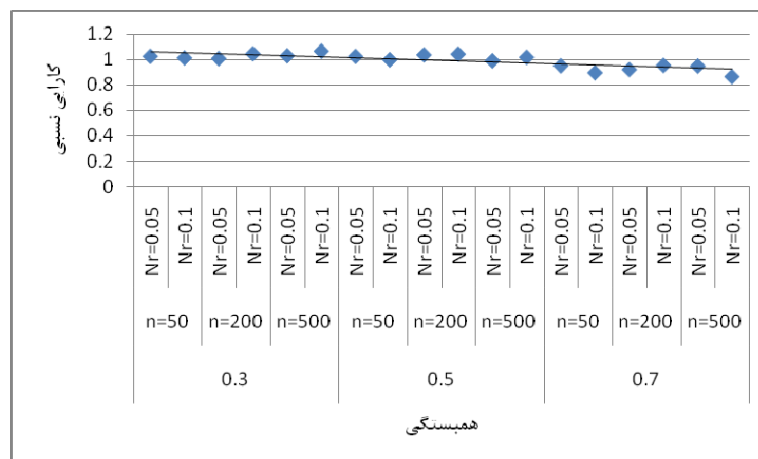
نمونه‌ی جدید	میانگین مشاهده‌های مشابه	الگوریتم EM	نرخ بی‌پاسخی	اندازه‌ی نمونه	همبستگی
۰/۰۱۳۴۰۳۱۵	۰/۰۵۶۳۴۲۱۴	۰/۰۱۵۴۰۱۴۶	۰/۰۵	۵۰	
۰/۰۱۳۵۲۰۶۴	۰/۰۱۸۵۶۱	۰/۰۱۵۸۶۶۳۸	۰/۱		
۰/۰۱۳۵۸۹۲۴	۰/۰۴۷۵۶۰۳۶	۰/۰۱۷۲۰۲۳۴	۰/۰۵	۲۰۰	۰/۳
۰/۰۱۲۶۲۵۸۸	۰/۰۸۰۴۳۸۶۵	۰/۰۱۸۲۰۴۰۴	۰/۱		
۰/۰۱۴۱۱۰۶۵	۰/۰۳۹۶۲۳۳۷	۰/۰۱۶۲۸۵۹	۰/۰۵	۵۰۰	
۰/۰۱۳۰۱۶۵۱	۰/۰۶۵۹۴۵۰۵	۰/۰۱۶۱۲۸۰۶	۰/۱		
۰/۰۰۰۸۱۰۱۲۲۵	۰/۰۴۴۱۳۲۵۳	۰/۰۰۳۴۱۲۳۲۷	۰/۰۵	۵۰	
۰/۰۰۲۹۲۳۱۸۵	۰/۰۹۰۱۳۰۰۴	۰/۰۰۳۱۳۳۹۲۱	۰/۱		
۰/۰۰۱۷۰۲۷	۰/۰۳۴۸۷۶۲۹	۰/۰۰۵۶۹۸۷۷۷	۰/۰۵	۲۰۰	۰/۵
۰/۰۰۱۴۷۳۷۰۷	۰/۰۶۹۱۵۴۰۴	۰/۰۰۷۶۶۳۱۹۴	۰/۱		
۰/۰۰۰۳۰۵۳	۰/۰۲۸۷۹۳۵۱	۰/۰۰۵۲۹۵۱	۰/۰۵	۵۰۰	
۰/۰۰۱۸۷۷۲۳	۰/۰۵۳۷۰۶۰۷	۰/۰۰۳۹۷۳۴۴۳	۰/۱		
۰/۰۰۴۶۶۰۶۲	۰/۰۴۷۴۲۵۵۲	۰/۰۰۶۴۸۰۱۸۳	۰/۰۵	۵۰	
۰/۰۰۲۱۲۵۵۰۲	۰/۰۹۱۴۴۷۳۷	۰/۰۰۴۹۱۸۶۱۴	۰/۱		
۰/۰۰۳۴۰۴۹	۰/۰۳۳۸۴۷۷۵	۰/۰۰۵۷۲۹۶۲۹	۰/۰۵	۲۰۰	۰/۷
۰/۰۰۲۴۴۶۵۱۶	۰/۰۶۸۳۵۵۲۴	۰/۰۰۹۵۴۱۰۴۶	۰/۱		
۰/۰۰۰۲۷۲۳۸۵۷	۰/۰۲۵۵۷۰۴۲	۰/۰۰۲۷۷۱۱۴	۰/۰۵	۵۰۰	
۰/۰۰۱۷۳۵۴۴۳	۰/۰۵۲۷۲۴۲۲	۰/۰۰۳۷۸۴۷۴۸	۰/۱		

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

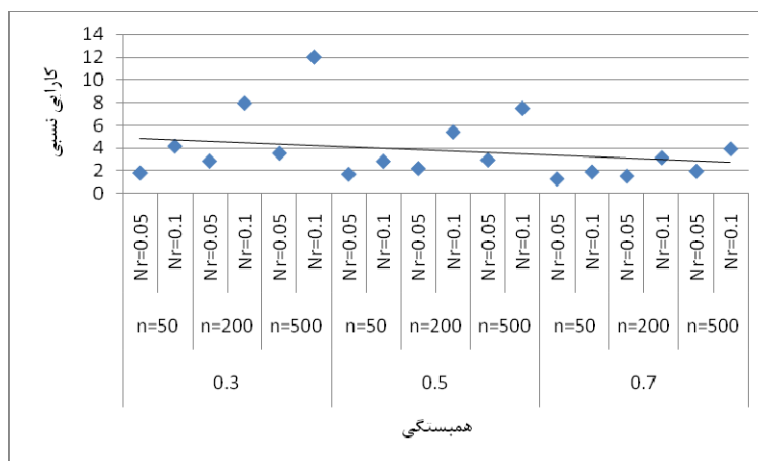
معیار کارایی نسبی در شکل ۲ نشان می‌دهد که دو روش جانهی با الگوریتم EM و جانهی با نمونه‌ی جدید وقتی که همبستگی بین دو دوره ۰/۳ و ۰/۵ باشد، تفاوت چندانی ندارند، در حالی که وقتی همبستگی بین دو دوره ۰/۷ باشد روش جانهی با الگوریتم EM بهتر از روش جانهی با نمونه‌ی جدید است.



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 کارا تر 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.



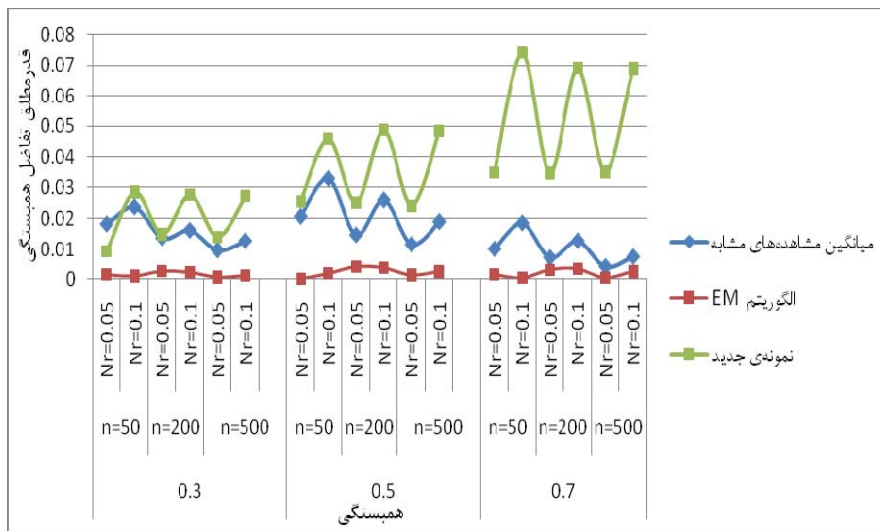
شکل ۳- کارایی نسبی روش جانهی با میانگین مشاهده‌های مشابه نسبت به نمونه‌ی جدید برای تمام حالت‌های مورد نظر

جدول ۲- مقایسه‌ی درستی پیشگویی سه روش جانهی

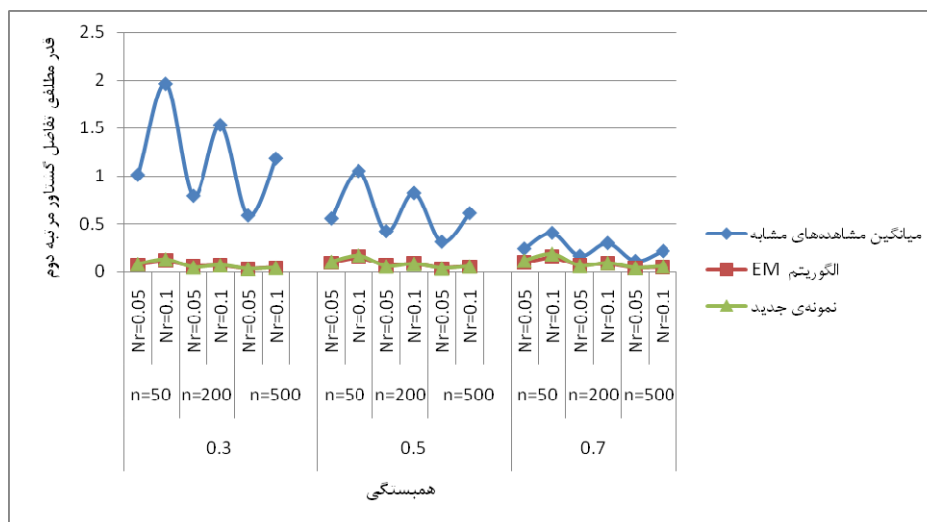
نمونه‌ی جدید	میانگین مشاهده‌های مشابه	الگوریتم EM	نرخ بی‌پاسخی	اندازه‌ی نمونه	همبستگی
۰/۳۴۷۰۵۳	۰/۸۰۲۵۸۴۵	۰/۹۴۸۲۰۰۶	۰/۰۵	۵۰	
۰/۱۷۲۷۹۱۶	۰/۶۷۴۵۵۲۹	۰/۹۰۷۲۴۸	۰/۱		
۰/۰۸۸۷۷۸۳۴	۰/۸۱۷۴۰۷	۰/۹۴۴۴۱۷	۰/۰۵	۲۰۰	۰/۳
۰/۰۳۶۲۴۵۶۵	۰/۶۹۹۹۰۱	۰/۸۹۹۲۸۲۵	۰/۱		
۰/۰۳۵۵۹۶۲۳	۰/۸۴۷۸۱۵۳	۰/۹۵۲۰۲۷۵	۰/۰۵	۵۰۰	
۰/۰۱۶۶۸۷۶۸	۰/۷۳۴۲۸۰۵	۰/۹۰۲۴۱۵۸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

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 برآورد 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 برآورد کم‌تر می‌شود.



شکل ۴- مقایسه‌ی درستی پیشگویی تغییر سه روش جانهی



شکل ۵- مقایسه‌ی درستی برآورد سه روش جانهی

لذا روش جانهی با الگوریتم EM از نظر معیار درستی برآورد در آمارگیری پانلی خانواری انگلیس نسبت به دو روش جانهی با میانگین مشاهده‌های مشابه و نمونه‌ی جدید عمل‌کرد بهتری دارد. میزان درستی برآورد 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 برآورد 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 برآورد و درستی پیشگویی تغییر روش جانهی با الگوریتم EM نیز نسبت به دو روش دیگر عمل‌کرد بهتری دارد (وقتی که برآورد 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 برآورد 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. and 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 رشیدی نژاد
کارشناس ارشد آمار
تهران، خیابان شهید بهشتی، نبش احمد قصیر، دانشکده‌ی اقتصاد دانشگاه علامه طباطبائی، گروه آمار.
رایانشانی: assiehrashidi@yahoo.com

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