مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۲، شماره‌ی ۱، بهار و تابستان ۱۳۹۰، صص ۴۵ – ۶۲

کاهش اریبی بی‌پاسخی توسط براوردگرهای کالبیده در مطالعه‌های پانلی

زهرا امینی فارسانی* و حمیدرضا نواب‌پور

دانشگاه علامه طباطبایی

چکیده: در بسیاری از نظام‌های آماری، نوعی از آمارگیری موسوم به آمارگیری در طول زمان متداول است که به‌شکلی مستمر در دوره‌های زمانی تکرار می‌شود. یکی از روش‌های آمارگیری در طول زمان، آمارگیری پانلی نام دارد. در این روش، به نمونه‌ای ثابت در دوره‌های زمانی مختلف مراجعه می‌شود. مهمترین چالش در یک آمارگیری پانلی، ناتوانی در به‌دست آوردن پاسخ از واحدهای نمونه‌ای و به‌وجود آمدن بی‌پاسخی است که معمولاً باعث اریبی و در برخی موارد منجر به افزایش واریانس براوردها می‌شود. در آمارگیری‌های پانلی علاوه بر بی‌پاسخی قلم اطلاعاتی، بی‌پاسخی دوره نیز وجود دارد. در این نوع بی‌پاسخی، واحد نمونه‌ای حد اقل برای یک دوره‌ی آمارگیری پاسخ‌گو است. حالت خاصی از بی‌پاسخی دوره، کاهش پاسخ‌گو است که در این حالت واحد نمونه‌ای از دوره‌ای به بعد از مطالعه خارج شده و هرگز به مطالعه باز نمی‌گردد. یکی از روش‌های کاهش اثرهای نامطلوب بی‌پاسخی، وزن‌دهی است. از جمله‌ی این روش‌ها، براوردگر کالبیده می‌باشد. در این مقاله پس از معرفی مفهوم‌های اولیه‌ی آمارگیری پانلی، انواع گم‌شدگی در آمارگیری‌های پانلی و ساختارهای گم‌شدگی، روش براوردگر کالبیده به‌عنوان روشی برای وزن‌دهی داده‌های گم‌شده معرفی می‌شود. سپس با استفاده از داده‌های طرح آمارگیری پانلی خانواری انگلستان، روش براوردگر کالبیده با روش براوردگر رگرسیونی تعمیم‌یافته از نظر معیارهای قدر مطلق اریبی نسبی، میانگین توان دوم خطا و کارایی نسبی مجانبی مقایسه می‌شوند. یافته‌های این مطالعه نشان می‌دهند که وقتی همبستگی بین دو دوره بالا بوده و ساختار گم‌شدگی تصادفی است، روش براوردگر کالبیده نسبت به سایر روش‌ها عمل‌کرد بهتری دارد.

---

* نویسنده‌ی عهده‌دار مکاتبات

دریافت: ۱۳۸۹/۲/۲۹، پذیرش: ۱۳۹۰/۶/۱۴.

واژگان کلیدی: آمارگیری پانلی؛ بی‌پاسخی دوره؛ گم‌شدگی تصادفی؛ وزن‌دهی، براوردگرهای کالبیده؛ براوردگرهای رگرسیونی تعمیم‌یافته.

۱- مقدمه

یک طرح آمارگیری عبارت است از تصمیم‌گیری برای انتخاب اعضای جامعه‌ای که مایلیم از اعضای آن اطلاعات آماری به دست آوریم. معمولاً طرح‌های نمونه‌گیری را می‌توان بر اساس تعداد مراحل نمونه‌گیری، نوع واحدهای گزینش‌شده در هر مرحله، چگونگی طبقه‌بندی واحدها قبل از انتخاب، روش انتخاب و تعداد واحدهای گزینش‌شده در هر مرحله از هم تمیز داد. ماهیت یک طرح نمونه‌گیری باید به گونه‌ای باشد که ویژگی‌های آماری براوردگرهای ارایه‌شده را برای خصیصه‌های مورد نظر بیان کند و بتواند با حفظ دقت در سطح مورد نظر، کم‌ترین هزینه را نیز در بر داشته باشد. باید توجه داشت که طرح‌های آمارگیری نمونه‌ای نسبت به مسائل مورد بررسی و با درنظر گرفتن هدف‌های طرح به حالت‌های مختلفی اجرا می‌شوند. در اجرای یک طرح آمارگیری دو گام اساسی وجود دارد: گام اول تعیین هدف‌ها و گام دوم شناخت محدودیت‌ها است. هدف‌های آمارگیری در واقع پرسش‌هایی هستند که در مورد جامعه‌ی مورد نظر پرسیده می‌شوند و محدودیت‌ها نیز در مورد مسائلی چون منابع مالی، نیروی انسانی، تمایل و توانایی پاسخ‌گویان مطرح می‌شوند. بی‌پاسخی و نقص چارچوب از عیب‌های هر طرح آمارگیری هستند. بدون این دو عامل مضر، کیفیت آماره‌ها و دقت براوردگرها بهبود می‌یابد. هیچ‌یک از این دو عامل را در مرحله‌ی طراحی آمارگیری نمی‌توان نادیده گرفت یا به‌طور کامل حذف کرد. بی‌پاسخی باعث کاهش دقت در آماره‌های آمارگیری می‌شود. در آمارگیری پانلی که یکی از مهم‌ترین انواع آمارگیری‌ها در طول زمان است، از آن‌جا که اندازه‌ی نمونه‌ای در طول دوره‌ها ثابت می‌باشد، بی‌پاسخی در آن‌ها یک مسئله‌ی جدی به‌شمار می‌آید. زیرا درحضور بی‌پاسخی، واریانس براوردگرها در هر دوره به علّت کاهش اندازه‌ی نمونه‌ای افزایش می‌یابد و براوردگرهای اریب نیز تولید می‌شوند. در حقیقت بی‌پاسخی عبارت است از عدم موفقیت در دریافت اطلاع مفید در مورد یک یا چند متغیر مورد بررسی برای یک یا چند عضو گزینش‌شده برای آمارگیری. این بی‌پاسخی‌ها به دلیل‌های مختلف رخ می‌دهند. مثلاً ممکن است پاسخ‌گو سؤال را به‌طور کامل متوجه نشده باشد، از آشکار شدن اطلاعاتش هراس داشته باشد و یا این که فرد بیمار باشد و

توانایی پاسخ‌گویی نداشته باشد. آگاهی از نوع بی‌پاسخی از اهمیت خاصی برخوردار است، چرا که نوع بی‌پاسخی بر روی انتخاب روشی که برای تعدیل آن به کار برده می‌شود، اثر می‌گذارد ([۶] و [۷]).

نکته‌ی قابل توجه در تعدیل اثر بی‌پاسخی، استفاده از اطلاعات کمکی است. این اطلاعات کمکی همان داده‌های ثبتی و یا اطلاعاتی هستند که از سرشماری‌های قبلی برای هر واحد در جامعه در دسترس می‌باشند. اطلاعات کمکی نه تنها در مرحله‌ی طراحی یک آمارگیری خوب به کار می‌روند بلکه با استفاده از این اطلاعات کمکی، نتیجه‌های مطلوب‌تری حاصل می‌شوند. بخصوص زمانی که همبستگی متغیر پاسخ و این متغیرهای کمکی قوی باشد، دقت براوردهای حاصل به مراتب بیش‌تر و اریبی ناشی از بی‌پاسخی تا حد ممکن کاهش می‌یابد. این‌که چه اطلاعات کمکی برای هر واحد نمونه‌ای انتخاب شود، حایز اهمیت است. در این مقاله روش‌هایی را معرفی می‌کنیم که با استفاده از اطلاعات کمکی که برای پاسخ‌گوها و بی‌پاسخ‌ها در دسترس هستند، اریبی ناشی از بی‌پاسخی واحد را تا حد ممکن تعدیل می‌کنند.

در بخش دوم این مقاله، به معرفی آمارگیری پانلی پرداخته می‌شود. در بخش سوم، ساختارهای گم‌شدگی به اختصار بیان شده و در بخش چهارم روش براوردگر کالبیده به‌عنوان روشی برای وزن‌دهی داده‌های گم‌شده معرفی می‌شود. در بخش پنجم با انجام یک مطالعه، اثر استفاده از دو روش وزن‌دهی براورد رگرسیونی تعمیم‌یافته و براورد کالبیده در تعدیل اثر بی‌پاسخی مقایسه می‌شوند.

۲- آمارگیری پانلی

پدیده‌ها و داده‌های اطراف ما درطول زمان پیوسته در حال تغییر می‌باشند. عقیده‌ها، نظرها و حالت‌های افراد در طول زمان تغییر می‌کنند (مانند تغییر عقیده‌های افراد در ارتباط با یک موضوع خاص، تغییر نرخ بیکاری و یا تورم در تحول‌های اقتصادی). جامعه‌شناسان، سیاست‌مداران و دیگر محققان به اطلاع از روند این تغییرها نیاز دارند تا بتوانند از آن‌ها برای تدوین برنامه‌ریزی‌ها و سیاست‌گذاری‌ها استفاده کنند. یک نامزد انتخاباتی برای براورد شانس پیروزی خود در انتخابات به روند تغییر نظرهای مردم نیازمند است و یا دانشمندان علوم‌پزشکی برای کشف راه‌های درمان یک بیماری به بررسی تغییرِ علائم بیماری در طول زمان و نرخ این تغییرها می‌پردازند. این‌گونه نیازهای

آماری، انگیزه‌ای برای در نظر گرفتن عامل زمان در بررسی‌ها و استفاده از آمارگیری مکرر شدند. البته تغییرِ ویژگی‌های جامعه و ترکیب آن‌ها، سادگی تحلیل طرح‌های مقطعی را از بین می‌برد و در عین حال باعث تنوع هدف‌های تحلیل و کاربردهای آن‌ها می‌شود.

میزان اهمیت آمارگیری مکرر را می‌توان در کاربردها و توانایی‌های این نوع آمارگیری، تنوع تحلیل‌های مربوط به آن و در برخی موارد کاهش هزینه و بالا بردن دقت براوردهای حاصل از آن خلاصه کرد. یکی از مزیت‌های داده‌های پانلی براوردِ تغییر پارامترهای مورد بررسی در طی زمان است- حتی اگر این تغییرها ناچیز باشند- مانند اندازه‌ی تغییر در نرخ بیکاری به‌صورت ماهانه.

آمارگیری پانلی از معدود روش‌های آمارگیری مکرر است که اجازه‌ی مطالعه‌ی تغییرها در سطح واحدهای کوچک را به پژوهش‌گر می‌دهد. همچنین برخی خطاهای حاصل در نمونه‌گیری ابتدایی (از قبیل بی‌پاسخی، عدم درج صحیح اطلاعات و...) در مراحل بعد قابل رفع‌شدن می‌باشند. به این ترتیب هزینه‌های پاکسازی داده‌ها بسیار کاهش می‌یابند. در آمارگیری پانلی، ثابت بودن نمونه در دوره‌های زمانی مختلف، به حذف «اثر نمونه» در شناسایی تغییرها کمک خواهد کرد. به‌عبارت دیگر تفاوت مشاهده‌شده در دو دوره‌ی زمانی را نمی‌توان ناشی از تفاوت واحدهای نمونه‌ای قلمداد کرد، لذا این نوع آمارگیری برای براورد تغییرها مفید است. لاوتون و پاس [۴] نشان دادند که آمارگیری پانلی در مقایسه با دیگر روش‌های آمارگیری مکرر می‌تواند تا ۵۰ درصد از هزینه‌ی آمارگیری در نوبت‌دوم بکاهد. زیرا یافتن و ارتباط با فردی که قبلاً با او مصاحبه شده است به مراتب ارزان‌تر از مصاحبه با یک فرد جدید می‌باشد. برای مثال در مطالعه‌ی پانلی پویایی درامد در آمریکا اولین مصاحبه به‌صورت ملاقات حضوری انجام شد، در حالی که مصاحبه در سایر دوره‌ها به‌صورت تلفنی صورت می‌گیرد.

در آمارگیری پانلی نوعی بی‌پاسخی وجود دارد که به‌عنوان بی‌پاسخی دوره شناخته می‌شود. اگر از واحدهای نمونه‌ای در یک یا چند دوره‌ی آمارگیری و نه در تمام دوره‌ها هیچ‌گونه پاسخی دریافت نشود، آن را بی‌پاسخی دوره گویند. کاهش پاسخ‌گو حالت خاصی از بی‌پاسخی دوره است، به این معنی که بعضی از واحدهای نمونه‌ای از دوره‌ای به بعد از مطالعه خارج شوند و تا انتهای مطالعه قابل دسترسی نباشند. لیتل و دیوید [۵]، سه نوع بی‌پاسخی دوره را معرفی کردند که شامل کاهش پاسخ‌گو (Attrition)، بازگشت (Reentry) و ورود با تأخیر (Late entry) می‌باشند.

بازگشت: زمانی رخ می‌دهد که یک واحد نمونه‌ای از دوره‌ای به بعد از مطالعه خارج شود و پس از آن دوباره مراجعه کند.

ورود با تأخیر: وقتی یک واحد نمونه‌ای در دوره‌ی اول در مطالعه حضور نداشته باشد و پس از آن وارد شود.

اگر بی‌پاسخی دوره به عنوان مجموعه‌ای از بی‌پاسخی‌های واحد قلمداد شود، وزن‌دهی، رهیافتی مناسب برای تعدیل اثر بی‌پاسخی است.

۳- ساختارهای گم‌شدگی

همان‌طور که اشاره شد، به دلیل این‌که در آمارگیری‌های پانلی امکان رخداد بی‌پاسخی دوره وجود دارد، اکنون به معرفی ساختارهای گم‌شدگی می‌پردازیم. متغیر نشانگر پاسخ R_{it} را به صورت زیر تعریف می‌کنیم:

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

فرض کنید که $Y_{i۱}^{o}$ نمایانگر مقدار مشاهده شده‌ی واحد iام در دور اول آمارگیری و $Y_{i۲}^{m}$ مقدار گم‌شده‌ی واحد iام در دور دوم آمارگیری باشد. احتمال گم‌شدگی واحد iام 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 [۸] 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

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

به‌عبارت دیگر احتمال این که پاسخ واحد iام در دور دوم گم‌شده باشد به مقدار پاسخ مشاهده‌شده‌ی واحد iام در دور اول یا مقدار پاسخ واحد iام در دور دوم که ممکن است گم‌شده باشد، بستگی ندارد. در حالتی که گم‌شدگی متغیر پاسخ در دور دوم به‌دلیل پاسخی که در دور اول مشاهده شده رخ دهد، گم‌شدگی تصادفی (Missing at Random) است. یعنی:

$$P(R_{i2}=0 \mid Y_{i1}^{o}, Y_{i2}^{m}) = P(R_{i2}=0 \mid Y_{i1}^{o}) = p_i^{o}.$$

به عبارت دیگر احتمال گم‌شدگی پاسخ واحد iام در دور دوم وابسته به مقدار مشاهده‌شده‌ی واحد iام در دور اول است. اگر احتمال مقادیر گم‌شده، به مقادیر مشاهده‌نشده‌ای که باید به دست می‌آید، وابسته باشد گم‌شدگی را غیر تصادفی (Not Missing at Random) نامید. در این حالت:

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

۴- براوردگرهای کالیبده

همان‌طور که اشاره شد، یکی از روش‌های کاهش اثر نامطلوب بی‌پاسخی دوره، استفاده از روش‌های وزن‌دهی است که از جمله‌ی این روش‌ها، روش براوردگرهای کالیبده می‌باشد. ابتدا پس از معرفی روش براوردگرهای کالیبده در حالت کلی (بدون بی‌پاسخی)، به معرفی روش براوردگرهای کالیبده به‌عنوان روشی برای وزن‌دهی داده‌های گم‌شده می‌پردازیم. جامعه‌ی متناهی $U=\{1,...,k,...,N\}$ شامل N واحد را در نظر بگیرید. این جامعه، جامعه‌ی هدف نام دارد. هدف ما براورد مجموع $t_y = \sum_U y_k$ است که y_k مقدار متغیر مورد بررسی، Y، برای k امین واحد جامعه‌ی هدف است. فرض کنید نمونه‌ی احتمالاتی s به اندازه‌ی n از جامعه‌ی هدف U و با احتمال $p(s)$ انتخاب شده است. احتمال‌های انتخاب معلوم و برای تمام $k \in U$ عبارت‌اند از $\pi_k = \sum_{k \in s} p(s)$ که به آن‌ها احتمال‌های شمول گفته می‌شود. فرض می‌شود طرح به‌گونه‌ای است که برای تمام واحدهای k داشته باشیم $\pi_k > 0$. مقدار $d_k = 1/\pi_k$ وزن طرح (design weight) برای واحد kام نامیده می‌شود. برای محاسبه‌ی براورد واریانس، به احتمال‌های شمول

مرتبه‌ی دوم نیاز است که با π_{kl} نشان داده می‌شوند. از طرفی π_{kl} عبارت است از احتمال این‌که واحدهای k و l هر دو در نمونه‌ی s وجود داشته باشند و از رابطه‌ی $\pi_{kl} = \sum_{\{k,l\}\in s} p(s)$ محاسبه می‌شود. وزن متناظر با آن عبارت است از $d_{kl} = 1/\pi_{kl}$. باید توجه داشت که $\pi_{kk} = \pi_k$ و $d_{kk} = d_k$.

در فرایند کالیبدن، براوردگرهای کالیبده از بردار اطلاعات کمکی X_k^* استفاده می‌کنند تا براوردگرهای کاراتری تولید کنند. کارایی براوردگرهای کالیبده بستگی به این دارد که آیا متغیرهای کمکی به‌خوبی، تغییرپذیری متغیر مورد مطالعه‌ی y را توضیح می‌دهند؟ علاوه بر ویژگی‌های بالا، وزن‌های کالیبیده‌ی w_k فاصله‌ی خی‌دوی $\sum_{k\in s} \frac{(w_k - d_k)^2}{d_k q_k}$ را نیز مینیمم می‌کنند که در آن q_k ها مقدارهای مثبت، معلوم و مستقل از d_k هستند. به‌طور کلی در همه‌ی براوردگرهای کالیبده به‌صورت $\hat{t}_{yw} = \sum_{k\in s} w_k y_k$، با وزن‌های w_k ای که در معادله‌های کالیبده صدق کرده و یک تابع فاصله را مینیمم می‌کنند، سر و کار داریم، [۲]، که در آن y_k مقدار صفت مورد نظر در k امین عضو نمونه‌ای و w_k وزن صحیح‌شده‌ی عضو k ام (وزن کالیبیده) است و عبارت است $w_k = d_k g_k$ که در آن d_k ها وزن‌های پایه (عکس احتمال انتخاب واحد k ام) و g_k تابعی از معیار فاصله‌ی انتخابی است ([۱] و [۶]). همان‌طور که اشاره شد، وزن‌های کالیبده باید میانگین تابع فاصله‌ی خی‌دو یعنی $E_P\left\{\sum_{k\in s} \frac{(w_k - d_k)^2}{d_k q_k}\right\}$ را مینیمم کنند که معادل با مینیمم کردن عبارت زیر است:

$$\sum_{k\in s} \frac{(w_k - d_k)^2}{d_k q_k}$$

با توجه به قید مقایسه‌ای (معادله‌ی کالیبیده) $\sum_{k\in s} d_k g_k X_k^* = \sum_U X_k^*$ و با استفاده از فن ضرایب نامعلوم لاگرانژ، عبارت بالا به‌صورت زیر مینیمم می‌شود. با مشتق گرفتن از عبارت زیر نسبت به w_k و مساوی صفر قرار دادن این عبارت، وزن‌های کالیبیده به‌صورت زیر حاصل می‌شوند:

$$\sum_s \frac{(w_k - d_k)^2}{d_k q_k} - 2\lambda \sum_s w_k X_k^*$$

$$w_k = d_k \left(1 + q_k X_k^{*\prime} \lambda\right)$$

ضریب λ نیز به‌صورت زیر به دست می‌آید:

$$\lambda = \left(t_x - \hat{t}_{x\pi}\right)\left(\sum_{k\in s} d_k q_k X_k^* X_k^{*\prime}\right)^{-1}$$

با جایگذاری λ در w_k و سپس جایگذاری w_k در $\sum_{k\in s} w_k y_k$، براوردگر رگرسیونی تعمیم‌یافته حاصل می‌شود:

$$\hat{t}_{y\mathrm{GREG}} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'\hat{B}$$

که در آن:

$$\hat{B} = \left(\sum_s d_k q_k X_k^* \left(X_k^*\right)'\right)^{-1} \left(\sum_s d_k q_k X_k^* y_k\right),$$

$$\hat{t}_{y\pi} = \sum_{k\in s} d_k y_k \text{ ،} \quad t_x = \sum_{k\in s} w_k X_k^* \text{ و } \hat{t}_{x\pi} = \sum_{k\in s} d_k X_k^*$$

است ([۳] و [۶]).

مشکلی که در محاسبه‌ی وزن‌های کالبیده وجود دارد این است که ممکن است وزن‌ها منفی شوند. تابع فاصله‌های مختلف، براوردگرهای کالبیده‌ی دیگری را نیز نتیجه می‌دهند که می‌توان به واسطه‌ی آن تابع‌های فاصله، مشکل منفی شدن وزن‌های حاصل از این روش را برطرف کرد [۲].

فاصله‌ی بین وزن‌های اصلی d_k و وزن‌های جدید w_k به‌طور دلخواه $(w_k - d_k)^2/(d_k q_k)$ در نظر گرفته شده بود. اما طبیعی است که اندازه‌های دیگری نیز برای این فاصله باید مورد بررسی قرار گیرند. لاندستروم و سارندال [۶] بیان می‌کنند براوردگرهای کالبیده در حالت پاسخ کامل توسط دویل و سارندال [۲] ارایه شد. آن‌ها به‌دنبال براوردگری به‌صورت $\hat{t}_{yw} = \sum_{k\in s} w_k y_k$ با وزن‌های w_k بودند، به‌طوری که این

وزن‌ها تا حد امکان به وزن‌های اصلی طرح، d_k، نزدیک باشند. این در حالی است که این وزن‌ها باید در معادله‌ی کالیبیده‌ی $\sum_{k\in S} w_k X_k^* = \sum_U X_k^*$ صدق کنند. در این قسمت قصد داریم از روش کالیبدن در حالتی استفاده کنیم که مقدارهای y_k فقط برای مجموعه‌ی پاسخ‌گوها ($r \le s$) معلوم باشد نه برای نمونه‌ی کامل. بنا بر این به‌دنبال وزن‌هایی هستیم که در معادله‌های کالیبیده‌ی

$$\sum_{k\in r} w_k X_k^* = \sum_{k\in s} d_k X_k^* \qquad (۱)$$

$$\sum_{k\in r} w_k X_k^* = \sum_U X_k^* \qquad (۲)$$

صدق کنند. تابع فاصله برای مینیمم کردن این وزن‌ها $\sum_{k\in r}\{(w_k - d_k)^۲/d_k q_k\}$ است. مقدارهای q_k مثبت هستند. زمانی که تمام واحدهای نمونه‌ای پاسخ‌گو باشند، $r = s$، این تابع فاصله منجر به تولید براوردگر رگرسیونی تعمیم‌یافته می‌شود ([۶]).

قضیه‌ی ۱. *بر اساس اطلاعات معلوم جامعه، مینیمم کردن تابع فاصله‌ی خی‌دو تحت معادله‌ی کالیبیده‌ی (۲) منجر به تولید براوردگر کالیبیده‌ی*

$$\hat{t}_{Uw} = \sum_{k\in r} d_k v_{Uk} y_k$$

می‌شود که در آن

$$v_{Uk} = ۱ + q_k \left(\sum_U X_k^* - \sum_{k\in r} d_k X_k^*\right)' \left(\sum_{k\in r} d_k q_k X_k^* X_k^{*\prime}\right)^{-۱} X_k^*$$

برای $k \in r$ برقرار است.

عبارت بالا را به‌صورت زیر می‌توان بازنویسی کرد:

$$v_{Uk} = ۱ + \lambda_r' X_k^*$$

در این عبارت v_{Uk} همان g_k (یکی از مؤلفه‌های وزن کالیبیده که در بخش قبل معرفی شد) است با این تفاوت که g_k در حالت پاسخ کامل تولید می‌شود و عبارت λ_r' عبارت است از:

$$\lambda_r' = \left(\sum_U X_k^* - \sum_{k\in r} d_k X_k^*\right)' \left(\sum_{k\in r} d_k q_k X_k^* X_k^{*\prime}\right)^{-1}$$

زمانی که $\sum_U X_k^*$ نامعلوم است، می‌توان از براوردگر ناااریب آن یعنی $\sum_s d_k X_k^*$ استفاده کرد. به این ترتیب براوردگر کالبیده به‌صورت زیر است:

$$\hat{t}_{sw} = \sum_{k\in r} d_k \nu_{sk} y_k$$

که در آن

$$\nu_{ks} = 1 + q_k \left(\sum_{k\in s} d_k X_k^* - \sum_{k\in r} d_k X_k^*\right)' \left(\sum_{k\in r} d_k q_k X_k^* X_k^{*\prime}\right)^{-1} X_k^*, \quad k \in r$$

برای تمام $k \in r$ برقرار است (اثبات در [۶]).

۵- یک کاربرد

۵-۱- طرح مطالعه

آمارگیری پانلی خانواری انگلیس (BHPS) 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. هدف اصلی اجرای این آمارگیری، بررسی تغییرهای اجتماعی و اقتصادی در سطح فرد و خانوار در انگلستان و شناسایی، مدل‌بندی و پیش‌بینی چنین تغییرهایی و بررسی علت‌های این تغییرها بر اساس متغیرهای اجتماعی و اقتصادی بوده است. این آمارگیری از سال ۱۹۹۱ تا کنون هر سال از ابتدای سپتامبر تا انتهای آوریل سال بعد اجرا می‌شود.

جامعه‌ی هدف این مطالعه، از داده‌های دور اول BHPS گرفته شد که اندازه‌ی آن ۱۰۲۶۴ فرد بالغ بالای ۱۶ سال است. هدف از انجام این مطالعه، مقایسه‌ی به کارگیری روش براوردگر کالبیده با روش براوردگر رگرسیونی تعمیم‌یافته برای تعدیل اثر بی‌پاسخی برای متغیر نرخ بیکاری با توجه به معیارهایی که در ادامه معرفی می‌شوند، است. طرح

مورد مطالعه دارای متغیرهای کمی و کیفی زیادی مانند وضعیت اشتغال، مدت زمان کار در هفته، زمان مسافرت، میزان پس‌انداز و ... می‌باشد. در این‌جا متغیر وضع فعالیت شامل واحدهای بیکار و شاغل، به‌عنوان متغیر کمکی که روش‌های وزن‌دهی با استفاده از آن اعمال می‌شوند، در نظر گرفته شده است. کدهای زیر به این متغیر اختصاص داده شد: کد ۱ به واحدهای جامعه‌ای بیکار و کد صفر به واحدهای جامعه‌ای شاغل. داده‌های دور دوم آمارگیری به‌منظور ساختن ضریب همبستگی‌های متفاوت به‌صورت زیر تولید شدند:

- مجموعه داده‌های ۱: ۹ درصد از واحدهای جامعه‌ای شاغل دور اول آمارگیری به‌تصادف به واحدهای بیکار و ۱۳ درصد از واحدهای جامعه‌ای بیکار در دور اول به شاغل تبدیل می‌شوند. این مجموعه داده‌ها همبستگی ۰/۷ با داده‌های دور اول آمارگیری ایجاد می‌کند؛
- مجموعه داده‌های ۲: ۶ درصد از واحدهای جامعه‌ای شاغل در دور اول آمارگیری به‌تصادف به واحدهای بیکار و ۱۰ درصد از واحدهای جامعه‌ای بیکار در دور اول به شاغل تبدیل می‌شوند. این مجموعه داده‌ها همبستگی ۰/۸ با داده‌های دور اول آمارگیری ایجاد می‌کند؛ و
- مجموعه داده‌های ۳: ۳درصد از واحدهای جامعه‌ای شاغل در دور اول آمارگیری به‌تصادف به واحدهای بیکار و ۷ درصد از واحدهای جامعه‌ای بیکار در دور اول به شاغل تبدیل می‌شوند. این مجموعه داده‌ها همبستگی ۰/۹ با داده‌های دور اول آمارگیری ایجاد می‌کند. همبستگی‌های ذکر شده همبستگی‌های کندال بین داده‌های دو دوره برای متغیر وضع فعالیت هستند.

از جامعه‌های تولید شده در قسمت قبل نمونه‌هایی به ترتیب به اندازه‌های ۱۰۰، ۲۰۰ و ۵۰۰ به صورت تصادفی ساده بدون جایگذاری انتخاب شدند. برای این نمونه‌ها در دور دوم آمارگیری به‌صورت تصادفی به ترتیب ۰/۰۵ و ۰/۱ (نرخ بی‌پاسخی) گم‌شدگی داده ایجاد شد. به‌منظور مقایسه‌ی اریبی بی‌پاسخی بر روی براوردهای آمارگیری و میانگین توان دوم خطا در دو روش وزن‌دهی، با ۱۰۰۰ بار نمونه‌گیری مستقل از داده‌های دور دوم و براورد میانگین نرخ متغیر پاسخ در دور دوم و واریانس آن، به‌عنوان براوردی از میانگین و واریانس توزیع نمونه‌گیری نرخ بیکاری محاسبه می‌شود. فرض می‌شود که تمام واحدهای نمونه‌ای در دور اول مطالعه حضور دارند و کاهش پاسخ‌گو در دور دوم به‌صورت

تصادفی رخ می‌دهد. براورد نرخ بیکاری از دو روش وزن‌دهی براورد کالبیده و براورد رگرسیونی تعمیم‌یافته وقتی که گم‌شدگی تصادفی در دور دوم آمارگیری رخ دهد، مقایسه می‌شوند.

روش وزن‌دهی خوب روشی است که در حضور بی‌پاسخی بتواند با انتخاب اطلاعات کمکی مناسب- اطلاعاتی که همبستگی زیادی با متغیر پاسخ دارند- براوردهایی با اریبی ناچیز و نزدیک به براوردهای حاصل از داده‌های کامل تولید کند. در ادامه، معیارهای ارزیابی برای مقایسه‌ی روش‌های وزن‌دهی مورد نظر ارائه می‌شوند.

۲-۵- معیارهای مقایسه

۱-۲-۵- قدرمطلق اریبی نسبی (Absolute Relative Bias)

برای براورد قدرمطلق اریبی نسبی عمل نمونه‌گیری در دور دوم آمارگیری برای حالت‌های مورد نظر (با استفاده از نرم‌افزار R) به تعداد ۱۰۰۰ بار اجرا می‌شود، (با همبستگی‌ها و اندازه‌های مورد نظر)، سپس نرخ‌های بی‌پاسخی روی آن اعمال شده و با استفاده از روش‌های وزن‌دهی مذکور اثر بی‌پاسخی تعدیل شده و نرخ و اریبی آن به‌صورت زیر براورد می‌شود:

$$\hat{\mathrm{Bias}}(\hat{p}_\rho) = \hat{E}(\hat{p}_\rho) - P_\rho = \frac{1}{1000}\sum_{h=1}^{1000}\hat{p}_{h\rho} - P_\rho = \bar{\hat{p}}_\rho - P_\rho$$

که در آن $\bar{\hat{p}}_\rho = (1/1000)\sum_{h=1}^{1000}\hat{p}_{h\rho}$ است. عبارت $\hat{p}_{h\rho} = (1/n)\sum_{i=1}^{n} w_{ih\rho} y_{ih\rho}$، براورد نرخ بیکاری برای نمونه‌ی n تایی h ام $(h = 1,\ldots,1000)$ استخراج‌شده از داده‌های دور دوم آمارگیری تولید شده با ضریب همبستگی ρ و P_ρ نرخ بیکاری در جامعه در دور دوم آمارگیری تولید شده با ضریب همبستگی ρ هستند. براورد قدر مطلق اریبی نسبی به‌صورت زیر محاسبه می‌شود:

$$\text{براورد قدر مطلق اریبی نسبی} = \frac{\left|\hat{\mathrm{Bias}}\left(\bar{p}_\rho\right)\right|}{\left(\hat{\mathrm{var}}(\hat{\mathrm{Bias}}(\hat{p}_\rho)\right)^{\frac{1}{2}}}$$

که در آن

$$\overline{\hat{\mathrm{Bias}}}(\hat{p}_{\rho}) = \frac{1}{1000}\sum_{h=1}^{1000}\hat{\mathrm{Bias}}(\hat{p}_{h\rho})$$

و $\hat{\mathrm{var}}(\hat{\mathrm{Bias}}(\hat{p}_{\rho}) = \frac{1}{1000-1}\sum_{h=1}^{1000}\left[\hat{\mathrm{Bias}}(\hat{p}_{h\rho}) - \overline{\hat{\mathrm{Bias}}}(\hat{p}_{\rho})\right]^2$ است.

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

اگر $\hat{\mathrm{var}}(\hat{p}_{\rho})$ براورد واریانس نرخ بیکاری برای همبستگی ρ باشد، آن‌گاه از رابطه‌ی زیر براورد میانگین توان دوم خطا به‌دست می‌آید:

$$\hat{\mathrm{MSE}}\left(\hat{p}_{\rho}\right) = \hat{\mathrm{var}}\left(\hat{p}_{\rho}\right) + \left[\hat{\mathrm{Bias}}\left(\hat{p}_{\rho}\right)\right]^2$$

براورد واریانس میانگین نرخ بر اساس ۱۰۰۰ تکرار نمونه‌گیری و از رابطه‌ی زیر بدست می‌آید:

$$\hat{\mathrm{var}}\left(\hat{p}_{\rho}\right) = \frac{1}{1000-1}\sum_{h=1}^{n}\left(\hat{p}_{h\rho} - \bar{\hat{p}}_{\rho}\right)^2$$

که در آن $\bar{\hat{p}}_{\rho} = \frac{1}{1000}\sum_{h=1}^{1000}\hat{p}_{h\rho}$ می‌باشد.

برای مقایسه‌ی دو روش وزن‌دهی، معیار براورد کارایی نسبی مجانبی از رابطه‌ی زیر محاسبه می‌شود:

$$\mathrm{EARE} = \frac{\mathrm{MSE}_{(\mathrm{cal})}}{\mathrm{MSE}_{(\mathrm{GREG})}}$$

این معیار با مقداری کم‌تر از یک دلالت بر کارا بودن روش کالیبدن نسبت به روش براورد رگرسیونی تعمیم‌یافته دارد.

۳-۵- یافته‌های مطالعه

پس از محاسبه‌ی براورد قدر مطلق اریبی نسبی، مشخص می‌شود که براورد قدر مطلق اریبی نسبی براورد نرخ بیکاری در دور دوم آمارگیری در روش براورد کالیبدن کم‌تر از

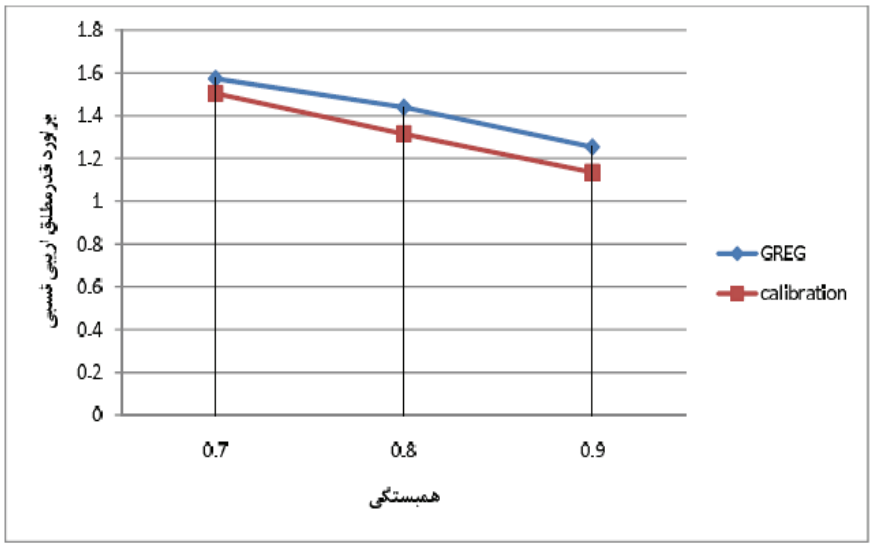
جدول ۱- قدر مطلق اریبی نسبی براورد نسبت بیکاری در دو روش وزن‌دهی با گم‌شدگی تصادفی

همبستگی	اندازه‌ی نمونه‌ای	کاهش پاسخ‌گو	براوردگر رگرسیونی تعمیم‌یافته	براوردگر کالبیده
۰/۷	۱۰۰	۰/۱	۱/۵۸۵۳	۱/۸۲۵۲۰۸
		۰/۰۵	۱/۵۷۳۴۶	۱/۵۰۳۶۱
	۲۰۰	۰/۱	۱/۳۳۴۰۰۲	۱/۳۲۳۵۶
		۰/۰۵	۱/۳۳۰۱۲۵	۱/۳۰۵۳۴
	۵۰۰	۰/۱	۱/۱۹۳۵۴	۱/۵۹۴۵۸
		۰/۰۵	۱/۱۳۴۱۷	۰/۵۹۹۳۴
۰/۸	۱۰۰	۰/۱	۱/۴۴۳۶۷	۱/۳۹۷۸۷۶
		۰/۰۵	۱/۴۳۹۳۴	۱/۳۱۴۹۳
	۲۰۰	۰/۱	۱/۲۹۲۳۴	۰/۷۰۳۴۵
		۰/۰۵	۱/۲۷۹۸۷۳	۰/۶۹۵۴۶۳
	۵۰۰	۰/۱	۰/۸۳۶۵۸۴	۰/۵۶۲۲۷
		۰/۰۵	۰/۸۱۸۳۴	۰/۵۴۲۵۵
۰/۹	۱۰۰	۰/۱	۱/۲۹۲۴۷	۱/۲۴۳۱۹
		۰/۰۵	۱/۲۵۴۲۶۷	۱/۱۳۴۴۲
	۲۰۰	۰/۱	۰/۸۴۳۵۶	۰/۶۶۵۴۸
		۰/۰۵	۰/۸۰۶۵۴	۰/۶۳۲۴۵
	۵۰۰	۰/۱	۰/۳۷۱۳۷	۰/۲۲۹۱۴
		۰/۰۵	۰/۱۷۴۶۹	۰/۰۰۱۵۴

براورد قدر مطلق اریبی نسبی روش براورد رگرسیونی تعمیم‌یافته است. در حالتی که همبستگی بین دو دوره و نرخ بی‌پاسخی کم باشد، روش براورد رگرسیونی تعمیم‌یافته نسبت به روش دیگر عملکرد خوبی دارد، اما با افزایش همبستگی بین دو دوره روش کالبیدن به مراتب بهتر عمل می‌کند (جدول ۱).

برای مثال در شکل ۱ با نرخ بی‌پاسخی $nr = ۰/۰۵$ و اندازه‌ی نمونه‌ای $n = ۱۰۰$، براورد قدر مطلق اریبی نسبی براورد نرخ بیکاری در روش کالیبدن از براورد قدر مطلق اریبی نسبی حاصل از روش براورد رگرسیونی تعمیم‌یافته کم‌تر است. در هر دو روش با افزایش همبستگی داده‌های دوره، براورد قدر مطلق اریبی نسبی براورد نرخ بیکاری کاهش می‌یابد. کم‌ترین مقدار براورد قدر مطلق اریبی نسبی در همبستگی ۰/۹ قابل مشاهده می‌باشد.

با ملاحظه‌ی اطلاعات جدول ۲، مربوط به براورد کارایی نسبی روش کالیبدن نسبت به روش براورد رگرسیونی تعمیم‌یافته در اندازه‌های نمونه‌ای ۱۰۰ و ۲۰۰ و ۵۰۰ و تمامی حالت‌های کاهش پاسخ‌گو، مشخص می‌شود که روش کالیبدن کاراتر از روش براورد رگرسیونی تعمیم‌یافته است، همچنین در شکل ۲ مشخص است با افزایش اندازه‌ی نمونه‌ای و افزایش نرخ بی‌پاسخی وزن‌دهی با استفاده از روش کالیبدن کاراتر می‌شود. در همبستگی ۰/۷ در دو مورد روش براورد رگرسیونی تعمیم‌یافته عمل‌کرد بهتری دارد، در حالی که وقتی همبستگی بین دو دوره ۰/۸ و ۰/۹ باشد روش کالیبدن بهتر از روش براورد رگرسیونی تعمیم‌یافته عمل می‌کند.



شکل ۱- براورد قدر مطلق اریبی نسبی برای نرخ بی‌پاسخی ۰/۰۵ و اندازه‌ی نمونه‌ای ۱۰۰

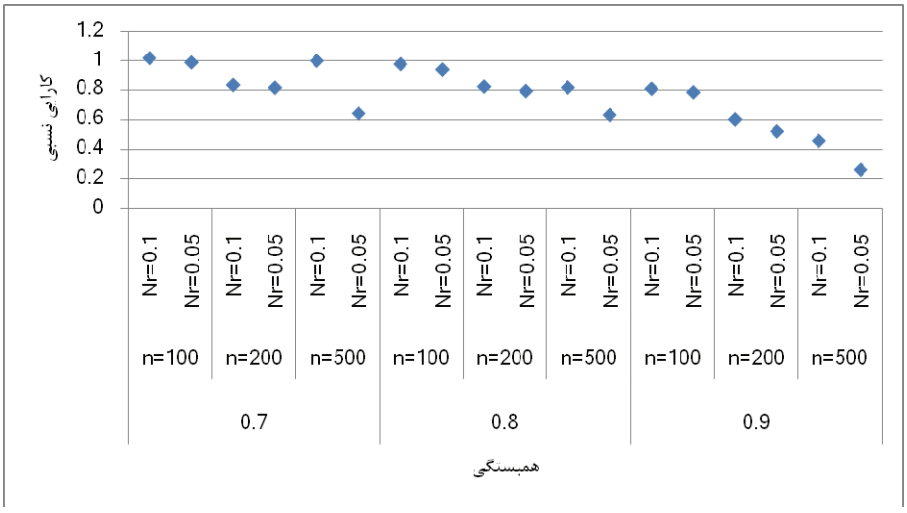
جدول ۲- براورد کارایی نسبی مجانبی روش براورد کالیبده نسبت به روش براورد رگرسیونی تعمیم‌یافته

همبستگی	$n=۱۰۰$		$n=۲۰۰$		$n=۵۰۰$	
	$nr=۰/۱$	$nr=۰/۰۵$	$nr=۰/۱$	$nr=۰/۰۵$	$nr=۰/۱$	$nr=۰/۰۵$
۰/۷	۱/۰۱۹۸	۰/۹۹	۰/۸۳۶۲	۰/۸۱۷۹	۱/۰۰۱	۰/۶۴۳۸
۰/۸	۰/۹۷۸۵	۰/۹۴۰۱	۰/۸۲۵۵	۰/۷۹۴۸	۰/۸۱۹۳	۰/۶۳۳۹
۰/۹	۰/۹۱۰۵	۰/۸۸۶۲	۰/۶۰۳۸	۰/۶۰۳۸	۰/۴۵۹۲	۰/۲۶۳۷

۶- خلاصه

با مطالعه‌ی شبیه‌سازی انجام‌شده، نتایج زیر حاصل می‌شود:
بر اساس معیارهای مقایسه می‌توان گفت: استفاده از روش کالیبدن برای داده‌های پانلی و براورد نرخ بیکاری و کاهش اثرهای نامطلوب بی‌پاسخی دوره با گم‌شدگی تصادفی (به‌جز در حالت $\rho=۰/۷$ ، $nr=۰/۱$ ، $n=۵۰۰،۱۰۰$) نسبت به روش براورد رگرسیونی تعمیم‌یافته، عمل‌کرد بهتری دارد.

برای اثر نرخ بی‌پاسخی، اندازه‌ی نمونه‌ای و همبستگی در براورد قدر مطلق اریبی نسبی برای هر دو روش کالیبدن و GREG می‌توان گفت:



شکل ۲-براورد کارایی نسبی مجانبی روش براورد کالیبده نسبت به روش براورد رگرسیونی تعمیم‌یافته

- با کاهش نرخ بی‌پاسخی براورد قدر مطلق اریبی نسبی در این روش کاهش می‌یابد،
- با افزایش اندازه‌ی نمونه‌ای، براورد قدر مطلق اریبی نسبی کاهش می‌یابد، و
- با افزایش همبستگی بین دوره‌ها، میزان براورد قدرمطلق اریبی نسبی کاهش می‌یابد.

این نتیجه‌ها برای هر دو روش برقرار است ولی برای روش کالیبدن کاهش‌های قدر مطلق اریبی نسبی بیش‌تر دیده می‌شود. از آن‌جا که همواره براورد قدر مطلق اریبی نسبی روش کالیبدن (به‌جز موردهای اشاره‌شده در جدول ۱)، کم‌تر از روش GREG است، بنا بر این استفاده از براوردگر کالیبده در حضور نرخ بی‌پاسخی کم و همبستگی قوی بین دوره‌ها و اندازه‌ی نمونه‌ای زیاد، عمل‌کرد خوبی برای براورد نرخ در آمارگیری‌های پانلی دارد.

نکته‌ی قابل توجه در استفاده از روش کالیبدن، کم بودن براورد قدر مطلق اریبی نسبی براورد نرخ و میانگین توان دوم خطای آن به‌خصوص در همبستگی ۰/۹ است که با توجه به زیاد بودن همبستگی بین دوره‌ها در اکثر آمارگیری‌های پانلی، ویژگی ارزشمندی برای این روش محسوب می‌شود. در واقع، افزایش همبستگی، کاهش اریبی و خطا را برای این روش به‌دنبال دارد.

استفاده از روش کالیبدن با توجه به اینکه از اطلاعات کمکی مناسب کمک گرفته و با داشتن وزن‌های پایه‌ای برای واحدهای پاسخ‌گو، وزن‌هایی تولید شود که بسیار به وزن‌های پایه‌ای طرح نزدیک شود، به تبع آن براوردهایی با اریبی کم در حضور بی‌پاسخی حاصل می‌شود، مناسب‌تر است. در واقع استفاده از این روش با توجه به کاهش اریبی با گم‌شدگی تصادفی و نیز تعدیل بهتر اثر بی‌پاسخی دوره نسبت به روش GREG، مثمر ثمرتر خواهد بود. استفاده از براورد قدر مطلق اریبی نسبی براوردهای تغییرهای نرخ بیکاری با استفاده از دو روش ذکر شده برای کارهای آتی پیشنهاد می‌شود.

مرجع‌ها

[1] Demnati, A. and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17-26.

[2] Deville, J.C. and Sarndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of American Statistical Association*, **87**, 376-382.

[3] Fuller, W.A. (2009). *Sampling Statistics*, First Edition, Wiley, Ames, Iowa.

[4] Lawton, T.K. and Pas, E.I. (1996). Resource Paper, Survey Methodologies Workshop, In Proceedings, Conference On Household Travel Surveys: New Concepts And Research Needs. Conference Proceedings Vol. 10, Transportation Research Board, Washington, Dc.

[5] Little, R.J.A. and David, M.H. (1983). Weighting adjusment for nonresponse in panel surveys". Working paper , U. S. Bureau of the Census, Washington, D.C.

[6] Lundstrom, S. and Sarndal, C.E. (2005). *Estimation in Survey with Nonresponse*, Third Edition, Wiley, New York.

[7] Moser, C.A. and Kalton, G. (1979). *Survey Methods In Social Investigation*, 2nd. ed. Heinemann Educational Books, London.

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

زهرا امینی فارسانی
فوق لیسانس آمار
تهران، خیابان شهید بهشتی، نبش احمد قصیر، دانشکده‌ی اقتصاد دانشگاه علامه طباطبایی، گروه آمار.
رایانشانی: zahraaminifarsani@yahoo.com

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