مجله گزیده‌مطالب آماری
سال ۱۸، شماره ۱، بهار و تابستان ۱۳۸۶، صص ۳۴-۲۱

یک روش استفاده از جانشین‌ها برای تعدیل بی‌پاسخی

مجتبی گنجعلی،[۱،*] حسن رنجی،[۲] نیما دانش‌پرور[۲]

[۱] دانشگاه شهید بهشتی
[۲] پژوهشکده آمار

چکیده. از راهکارهای رویارویی با بی‌پاسخی واحد در آمارگیری‌ها، جانشین‌سازی هر واحد بی‌پاسخ با یک واحد جدید است که جانشینی (substitute) واحد آماری نامیده می‌شود. واحدهای جانشین می‌تواند به‌طور تصادفی یا جورشده انتخاب شود. یکی از انتقاداتی که بر روش جانشین‌سازی وارد است آن است که چون در این روش واحدهای جانشین، پاسخگویان هستند ممکن است به‌طور راژمند (سیستماتیک) متفاوت از افرادی باشند که جانشین آنان شده‌اند و در نتیجه ممکن است نتایج آمارگیری اریب شوند. در این مقاله، مروری بر استفاده از روش جانشین‌سازی، انتقادات وارد بر آن و پیشنهادهای استفاده صحیح از این روش به‌گونه‌ای که حتی در حالت گم‌شدگی (بی‌پاسخی) غیر قابل چشم‌پوشی نیز مفید باشد، آورده شده است.

پرتال جامع علوم انسانی

۱- مقدمه

بی‌پاسخی واحد نمونه‌گیری وقتی روی می‌دهد که واحد نمونه‌گیری‌شده، هیچ اطلاعی در اختیار آمارگیر قرار ندهد ([۹] و [۱۸]). یکی از رهیافت‌های رویارویی با بی‌پاسخی واحد، جانشین‌سازی هر واحد بی‌پاسخ با یک واحد جورشده جدید است. با واحدهای جانشین وقتی تماس گرفته می‌شود که پاسخی برای واحد آمارگیری‌شده اولیه به دست نیامده

واژگان کلیدی: جانشینی؛ بی‌پاسخی واحد؛ تعدیل بی‌پاسخی؛ بی‌پاسخی غیر قابل چشم‌پوشی؛ جانهی چندگانه.

* نویسنده عهده‌دار مکاتبات

باشد. به هر حال چون جانشین‌ها پاسخگویان هستند ممکن است به‌طور راژمند از افـرادی که جانشین آن‌ها شده‌اند متفاوت باشند بـه ایـن معنـی کـه ممکـن اسـت مقـادیری کـه جانشین‌ها گزارش می‌دهند از مقادیری که باید برای بی‌پاسخ‌ها ثبـت مـی‌شـدند متفـاوت باشند و در نتیجه ممکن است نتایج آمـارگیری اریـب شـوند. روبـین و زانوتـو [۳۰] روش جدیدی معرفی کرده‌اند که جانشین‌های جورشده به‌گونه‌ای انتخاب می‌شوند که تـا آن‌جـا که ممکن است مشابه افراد اولیه باشند و امکان استفاده از جانهی چندگانـه بـرای جـانهی داده‌های گم‌شده فراهم آید. این روش تفاوت‌های راژمند بین بی‌پاسخ‌هـا و جانـشین‌هـا را تعدیل کرده و همچنین بی‌پاسخی غیر قابل چـشم‌پوشـی را تـصحیح مـی‌کنـد. یـادآوری می‌شود که بی‌پاسخی غیر قابل چشم‌پوشی به این معناست که سازوکار گـم‌شـدن داده‌هـا وابسته به پاسخ‌های مشاهده نشده است [۲۳].

روش‌های تعدیل بی‌پاسخی غیر قابل چشم‌پوشی، اغلب مـستلزم فـرض‌هـای قـوی درباره توزیع مشاهدات یا تحلیل حساسیت است که در آن با تغییر فرض‌های اساسی مدل، تغییر در برآورد پارامترها را بررسـی مـی‌کننـد ([۳]، [۱۲]، [۱۵]، [۲۵]، [۲۶]، [۲۷]، [۲۸] و [۲۹]). روش روبین و زانوتـو [۳۰] بـر طـرح نمونـه‌گیـری متمرکـز اسـت و هماننـد [۱۴] داده‌های اضافی (additional) را در واکنش به بی‌پاسـخی گـردآوری مـی‌کنـد. گلـین و دیگران [۱۴] فرض مـی‌کننـد کـه پاسـخگویانِ پیگیـری‌شـده (follow-up)، نمونـه‌ای تصادفی از بی‌پاسخ‌های اولیه‌اند ولی روبین و زانوتو [۳۰] داده‌های اضافی را از پاسخگویان جانشین به دست می‌آورند. روش جانشین سازی، زمانی مفید است که ساختاری خوشـه‌ای در جامعه برقرار باشد که لازم نیست قابل مشاهده باشد و خوشه‌ها مشخصه‌هایی دارند که هم به بی‌پاسخ‌ها و هم به پیشامدها وابسته‌اند. در ایـن حالـت روش روبـین و زانوتـو [۳۰] برآوردهایی به دست می‌دهنـد کـه نـسبت بـه برآوردهـای حاصـل از روش‌هـای تعـدیل بی‌پاسخی دیگر کم‌تر اریب‌اند.

در بخش بعد پژوهش‌های موجود درباره‌ی جانشینی بی‌پاسخی را مرور خواهیم کـرد. سپس در بخش ۳ روش روبین و زانوتو [۳۰] و ارزیابی آن‌ها (به‌وسیله شـبیه‌سـازی) آورده می‌شود. در انتها نتیجه‌گیری و پیشنهادها ارائه خواهند شد.

۲- روش‌های جانشینی برای تعدیل بی‌پاسخی

واحدهای جانشین آمارگیری که جانشین‌های میدانی نیز نامیده می‌شوند ([۶] و [۳۷]) واحدهای پاسخگوی اضافی‌اند که یا به‌صورت تصادفی بر حسب یک طرح نمونه‌گیری احتمالاتی (که معمولاً طرح نمونه‌گیری اولیه است) انتخاب می‌شوند یا طوری انتخاب می‌شوند که با پاسخگویان اولیه در یک یا چند مشخصه جور باشند. حالت اول را جانشین تصادفی و حالت دوم را جانشینی جورشده می‌نامند. واحدهای جانشین جای واحدهای بی‌پاسخ را گرفته و مجموعه داده‌های جدید، همانند یک مجموعه از داده‌های کامل تحلیل می‌شوند.

روشی مرتبط با این روش وجود دارد که گونه‌ای از جانشینی محسوب نمی‌شود و نمونه‌گیری مکمل (نمونه اضافی) نامیده می‌شود [۱۹]. این روش اندازه نمونه را با یک پیش‌بینی از نرخ بی‌پاسخی، به‌حدّی بالا می‌برد که در نهایت تعداد پاسخگوی مطلوب به دست آید. سپس برای تعدیل بی‌پاسخی، داده‌های پاسخگویان باز وزن‌دهی (reweight) می‌شوند یا داده‌های گم‌شده جانهی می‌شوند [۱۸]. به‌علت تشابه این روش با جانشینی تصادفی از این نظر که در هر دو روش برای جانشینی از طرح نمونه‌گیری اولیه استفاده می‌شود، نمونه‌گیری مکمل گاهی به‌عنوان یک روش جانشینی محسوب می‌شود.

در آمارگیری‌های متعددی جانشینی برای واحدهای بی‌پاسخ به‌کار رفته است که از میان آن‌ها می‌توان مطالعه طولی ملی (National Longitudinal Study) آمریکا ([۶] و [۳۹])، آمارگیری میشیگان از مصرف مواد (The Michigan Survey of Substance Use) [۳۳]، آمارگیری‌های انجام‌شده توسط وستات (Westat) [۳۸] و برخی آمارگیری‌های رسمی و دانشگاهی در اروپا [۳۷] را نام برد. مزیت اصلی جانشین‌سازی این است که می‌تواند نواحی به‌خصوصی در جامعه را که نرخ پاسخ‌دهی کمی دارند، هدف قرار دهد، و از این طریق کمک می‌کند که بخش‌های مختلف جامعه به‌طور مناسبی نمایانده شوند. این جنبه برای طرح‌های پیچیده بسیار مهم است به‌خصوص وقتی که در نهایت تعداد کمی از واحدها از هر طبقه یا خوشه انتخاب می‌شوند ([۶]، [۱۹]، [۲۲]، و [۳۷]).

بیش‌ترین انتقاد به روش جانشین‌سازی این است که این روش اریبی بی‌پاسخی را از بین نمی‌برد ([۵]، [۶]، [۱۳]، [۱۹]، [۲۲]، [۲۴]، و [۳۷]). به هر حال این انتقادی است که

به همه روش‌های تعدیل بی‌پاسخی وارد است (همین‌طور به روش‌هایی که تنها از داده‌های کامل استفاده می‌کنند). یادآور می‌شویم لسلر و کالسبیک [۲۲] بیان کرده‌اند که جانشینی میدانی که از جانشینی جورشده استفاده می‌کند اریبی بی‌پاسخی را کاهش می‌دهد. این امر به‌همان دلیلی رخ می‌دهد که روش‌های وزن‌دهی و جانهی بی‌درنگ یعنی جانهی به شرط متغیرهای مشاهده‌شده که سعی می‌کند پیوند بین متغیرهای مشاهده‌شده و گم‌شده را حفظ کند، اریبی را کاهش می‌دهند.

روش‌های جانشینی معایب دیگری نیز دارند. برای مثال، ممکن است در صورت اطلاع از استفاده از روش جانشینی برای واحدهای بی‌پاسخ، تلاش پرسشگر برای به دست آوردن اطلاعات از واحدهای نمونه‌گیری اولیه، کاهش یابد ([۶]، [۷]، [۱۷] و [۲۲]). به هر حال، سودمن [۳۴] اشاره کرده است که این مشکل لزوماً روی نمی‌دهد و چپمن و رومن [۷] ملاحظه کرده‌اند که این مشکل در نمونه‌هایی که در نظام مصاحبه تلفنی رایانه‌یار (computer-assisted) انتخاب شده‌اند، روی نداده است. دلیل، این است که در نظام مصاحبه تلفنی شیوه مصاحبه برای جانشین‌ها همانند آن برای واحدهای اولیه است. انتقاد دیگر به‌روش جانشینی آن است که اگر جانشین‌ها بعد از تلاش زیادی که برای جلب واحد اولیه انتخاب‌شده به‌کار رفته، انتخاب شوند، استفاده از آن ممکن است کار میدانی آمارگیری را طولانی کند ([۱۷]، [۳۷] و [۳۸]). در آن‌گونه مطالعات پزشکی که در آن‌ها واحدهای واجد شرایط در طی زمان انتخاب می‌شوند، چنین تأخیرهایی روی نمی‌دهند زیرا در این موقعیت‌ها جستجو برای انتخاب جانشین برای امتناع‌های زودرس (بی‌پاسخی واحد آماری بر اثر امتناع در تماس اولیه) با جستجو برای واحدهای آمارگیری اولیه تطابق دارند.

موضوع جانشین‌های آمارگیری به‌ندرت مطرح شده است. کتاب‌های درسی نمونه‌گیری خیلی مختصر به جانشین‌ها اشاره کرده‌اند ([۱۳]، [۱۹]، [۲۰] و [۲۲]) و برخی از آن‌ها اصلاً درباره آن مطلبی بیان نکرده‌اند ([۸]، [۱۶] و [۳۱]). کیش و هِس [۲۱] از اولین کسانی بوده‌اند که روش‌های جانشینی را مطرح کرده‌اند. به هر حال، کیش [۱۹] پیشنهاد می‌کند که از روش‌های جانشینی استفاده نشود و دلیل او آن است که جانشین‌ها شبیه پاسخگویان عمل می‌کنند، و نه شبیه بی‌پاسخ‌ها.

چپمن [۶] در یک مقاله مروری درباره استفاده از روش جانشینی، تنها چهار مطالعه

تجربی در مورد روش‌های جانشینی یافت ([۱۰]، [۱۱]، [۳۳] و [۳۹]). این مطالعات تلاش کرده‌اند تا اریبی ناشی از یک یا چند روش جانشینی را مورد بررسی قرار دهند. کار دوربین و استوارت [۱۱] تنها تحقیقی است که اریبی ناشی از روش‌های جانشینی و روش‌های دیگر جانهی را مقایسه می‌کند. در این مطالعه تفاوت کمی بین برآوردهای کلیدی که بر اساس مراجعات دوباره (callbacks) نامحدود به بی‌پاسخ‌ها به دست می‌آیند و برآوردهای به دست آمده از بهترین روش جانشینی (جانشین‌سازی بعد از ۳ بار تماس بدون موفقیت) وجود دارد. در این مطالعه نویسندگان دریافته‌اند که جانشین‌ها جوان‌تر از کسانی هستند که بعد از ۳ بار تماس گرفتن پاسخگو شده‌اند.

مطالعه کوهن [۱۰] و سیرکن [۳۳] نشان داد که بین جانشین‌ها و بی‌پاسخ‌ها تفاوت‌های معنی‌داری وجود دارند. به هر حال، این تفاوت‌ها ممکن است به‌دلیل تفاوت‌های بین پاسخگویان سریع و پاسخگویان کند (پاسخگویانی که دوباره به آنان مراجعه شده است) باشند. کوهن [۸] جانشینی را در صورتی به کار برده است که واحد اولیه در یک بار مراجعه پاسخگو نبوده است. سیرکن [۳۳] تنها یک بار به جانشین اولیه مراجعه کرده و در صورت عدم پاسخ به جانشین بعدی رجوع کرده است.

ویلیامز و فولسام [۳۹] تفاوت‌های معنی‌داری میان برآوردهای حاصل از به‌کارگیری روش جانشینی و وزن‌دهی، و برآوردهایی به دست آمده با استفاده از داده‌های حاصل از بی‌پاسخ‌های پیگیری‌شده نتیجه شده‌اند، به دست آورده‌اند. متأسفانه تأثیر به‌کارگیری جانشین‌ها از تأثیر تعدیل‌های وزنی بی‌پاسخی تفکیک نشده است.

بی‌مر و دیگران [۴] استفاده از جانشین‌ها را در یک آمارگیری با شماره‌گیری تلفنی تصادفی (Random Digit Dialing: RDD) بررسی کرده‌اند. در روش آنان جانشین برای هر بی‌پاسخ از داخل واحد نمونه‌گیری اولیه‌ای که بی‌پاسخ به آن تعلق داشته انتخاب شده است. آنان دریافتند که برآورد واریانس برای برآوردگر مبتنی بر جانشینی کوچک‌تر از برآورد واریانس برای برآوردگر مبتنی بر تعدیل وزنی است که از یک نمونه دیگر با هزینه‌ای برابر با هزینه روش جانشینی به دست آمده است. به هر حال، یک نمونه از بی‌پاسخ‌های پیگیری‌شده نشان داد به‌طور متوسط نسبت زنان و افراد مسن در خانوارهای جانشین‌شده بیش‌تر از بی‌پاسخ‌های متناظرشان می‌باشد. این موضوع ممکن است به‌دلیل

محدودیت زمانی موجود در مرحله‌ی گردآوری اطلاعات جانشین‌ها باشد که ایجاب می‌کند جانشین‌ها سریع‌تر از پاسخگویان کند نمونه‌ی اصلی باشند.

چپمن و رومن [۴] در مطالعه‌شان در استفاده از جانشین‌ها در یک آمارگیری RDD، نتایج مشابهی را گزارش کرده‌اند. وهووار [۳۷] گزارش می‌دهد که جانشین‌ها اریبی قابل توجهی را در آمارگیری نیروی کار اسلوواکی و آمارگیری اجتماعی اسلوواکی نتیجه داده‌اند. او گزارش مشابهی در مورد آمارگیری از هزینه خانوارهای ایتالیایی با استفاده از جانشین‌ها دارد.

به‌طور خلاصه تحقیق‌های تجربی نشان می‌دهند که جانشین‌ها (هرچند سریع) گزینه کاملاً مناسبی برای بی‌پاسخ‌ها نیستند. به‌علاوه چون جانشین‌ها پاسخگویان سریع را شبیه‌سازی می‌کنند روش‌های جانشینی ممکن است اریبی بیش‌تری از دیگر روش‌های تعدیل بی‌پاسخی همچون تعدیل‌های وزنی به دست دهند.

ناتان [۲۴] و وهووار [۳۷] استفاده از جانشین‌ها را از یک جنبه تحلیلی‌تر مورد بررسی قرار داده‌اند. ناتان [۲۴] برای برآورد نسبت جامعه یک روش افزایش اندازه نمونه (sample size inflation) را با یک روش جانشینی میدانی تصادفی مقایسه کرده است. او نشان داد که هر دوی این روش‌ها برای رسیدن به دقت یکسان، به تعداد نمونه و هزینه مورد انتظار یکسانی نیاز دارند. به هر حال، روش جانشینی، کنترل هزینه بهتری را نشان داد و متکی به برآورد نرخ بی‌پاسخی در مرحله طراحی نمونه‌گیری نبود.

وهووار [۳۷] با استفاده از یک نمونه‌گیری خوشه‌ای دو مرحله‌ای با جانشینی میدانی تصادفی و روش تعداد نمونه اضافی که با روش تعدیل وزنی ترکیب شده بود، مقایسه‌ای تحلیلی و تجربی از برآوردگر واریانس میانگین جامعه انجام داد. روش جانشینی، هر یک از بی‌پاسخ‌ها را با یک جانشین که از خوشه همانندش انتخاب می‌شد، یکسان جانشین می‌کرد. روش افزایش اندازه نمونه، تعداد نمونه اولیه را به‌منظور رسیدن به اندازه نمونه مورد انتظار افزایش می‌دهد، سپس وزن‌های نمونه‌ای پاسخگویان داخل یک خوشه را به‌گونه‌ای تعدیل می‌کند که متناسب با عکس نرخ پاسخ برآوردشده در آن خوشه باشند. با فرض آن که در داخل هر خوشه بی‌پاسخی کاملاً تصادفی داشته باشیم [۲۳]، اریبی پاسخ وجود نخواهد داشت. وهووار نشان داد که افزایش دقت به دست آمده از روش جانشینی

میدانی تصادفی نسبتاً کوچک است مگر برای موقعیت‌هایی که در آن‌ها نرخ بی‌پاسخی بالا بوده و تعداد نمونه داخل خوشه‌ها کوچک باشد. بر اساس ارزیابی‌های تجربی و فرض آن که محدودیت زمانی موجب می‌شود تا جانشین‌ها، پاسخگویانی سریع باشند، وهووار نتیجه گرفت که اریبی زیاد روش جانشینی (در مقایسه با روش افزایش اندازه نمونه که با تعدیل‌های وزنی ترکیب شده باشد) بر افزایش ناچیز دقت غلبه می‌کند. بنا بر این ظاهراً پژوهش‌های تحلیلی موجود محدود به مقایسه جانشینی تصادفی و روش افزایش اندازه نمونه است و این پژوهش‌ها حاکی از آن‌اند که جانشینی تصادفی، کارایی اندکی را که با اریبی زیاد همراه است در بر دارد.

تیترینگتون و سدرانسک [۳۵] یک روش جانهی را که در آن هر پاسخ گم‌شده y، با پاسخ اهداشده‌ای (donated) که در متغیر تبیینی پیوسته x با فرد اولیه جور است، ارزیابی کرده‌اند. یک نتیجه کلیدی شیبه‌سازی آنان، بر حسب قدر مطلق اریبی مورد انتظار و میانگین توان دوم خطای برآورد میانگین جامعه، آن است که جابه‌جایی پاسخ‌های گم‌شده با پاسخ‌های اهداشده خوب کار می‌کند، به شرط آن که اهداکنندگان با افراد اولیه بر حسب متغیر تبیینی خوب جور شده باشند. اگر جورسازی به‌خوبی انجام نشود (همان طور که در جانشینی آمارگیری‌ها ممکن است روی دهد چون جانشین‌ها اغلب پاسخگویان سریع‌اند)، در این صورت یک مدل (مانند رگرسیون خطی) که y را بر مبنای x جانهی می‌کند، جانهی را بهبود می‌بخشد (به‌خصوص اگر این مدل با استفاده از پاسخ‌هایی که به‌عنوان اهداکننده‌های جانهی انتخاب شده بودند، برازانده شود).

هر چند این پژوهش‌ها تأکید می‌کنند که از داده‌های جانشین نباید به‌جای داده‌های واحدهای بی‌پاسخ استفاده شوند، نمی‌توان از آن‌ها استنباط کرد که جانشینی کاملاً بی‌فایده است. آن‌قدر بدیهی نیست اگر داده‌هایی از جانشین‌ها موجود باشند که بتوان اطلاعات آن‌ها را نادیده گرفت و از روش‌های تعدیل سنتی همچون روش‌های وزن‌دهی و جانهی استفاده کرد ([۱۸] و [۲۹]). به‌خصوص که ممکن است بی‌پاسخ‌ها و جانشین‌ها شباهت‌هایی مهم داشته باشند که با مدل‌بندی‌های آماری بتوان از آن‌ها برای جانهی بی‌پاسخ‌ها استفاده کرد. هدف روبین و زانوتو که در بخش بعد می‌آید نیز چنین بوده است.

۳- جانهی چندگانه با استفاده از جانشینی در آمارگیری‌ها

روبین و زانوتو [۳۰] روش خود را جورسازی، مدل‌بندی و جانهی چندگانه (Matching, Modeling and Multiple imputation: MMM) نامیده‌اند. این روش از جانشین‌های جورشده برای جانهی داده‌های گم‌شده استفاده می‌کند. این شیوه، جانشین‌ها و بی‌پاسخ‌ها را طوری جورسازی می‌کند که بر حسب برخی متغیرهای کمکی که قبل از آمارگیری موجودند و برای جورسازی مناسب‌اند شبیه باشند (جورسازی متغیرهای کمکی). در نتیجه این جورسازی، جانشین‌ها و بی‌پاسخ‌ها ممکن است مقادیر مشابهی بر حسب متغیرهای کمکی که مشاهده شده‌اند، داشته باشند (متغیرهای کمکی میدانی). برای مثال اگر همسایه‌ها برای خانوارهای بی‌پاسخ جانشین شوند، ممکن است مشخصات مشابهی از نظر اقتصادی اجتماعی، سطوح رویارویی با حوادث غیر مترقبه، یا استفاده از وسایل نقلیه عمومی داشته باشند (هر چند که ممکن است این مشخصات ثبت نشده باشند).

روبین و زانوتو [۳۰] به‌جای روش معمول در استفاده از جانشین برای جانشینی بی‌پاسخ‌ها، آن‌ها را با توجه به اطلاعات موجود در مورد پاسخگویان و متغیرهای کمکی که برای هر دوی بی‌پاسخ‌ها و جانشین‌ها موجود است (مدل‌بندی متغیرهای کمکی) استفاده کردند تا مدلی برای جانهی چندگانه بی‌پاسخی فراهم شود. برای کمک در برازش این مدل برای پاسخگویانی که شبیه بی‌پاسخ‌ها بودند نیز انتخاب شدند. به‌محض آن که روی داده‌های بی‌پاسخ‌ها جانهی چندگانه انجام شد تمام داده‌های جانشین‌ها (برای هر دوی پاسخگویان و بی‌پاسخ‌ها) کنار گذاشته می‌شود. استفاده از جانهی چندگانه [۲۹] برآوردهای نقطه‌ای و برآورد واریانس برآوردگرها را حتی برای آمارگیری‌های پیچیده و پارامترهای مشکل جامعه، تسهیل می‌کند ([۱] و [۳۲]). در زیر مثال ساده‌ای از این روش تشریح می‌شود. در عمل مدل‌های پیچیده‌تری مورد نیاز است ولی مثال زیر ایده‌های اصلی را می‌شکافد.

فرض کنید رابطه زیر در جامعه برقرار باشد:

$$y_i = \beta_0 + \beta_1 x_{m,i} + \beta_2 x_{f,i} + \varepsilon_i \qquad (۱)$$

که در آن x_m متغیر تبیینی است که در مدل وارد شده است، x_f متغیر تبیینی میدانی و ε خطای تصادفی با $\varepsilon_i \sim N(0,\sigma^2)$ است. متغیر تبیینی مدل متغیری است که می‌تواند در مدل آماری برای تصحیح تفاوت‌های مشاهده‌شده بین بی‌پاسخ‌ها و جانشین‌ها وارد شده باشد، اما ممکن است برای جورسازی موجود نباشد یا استفاده نشود. برای مثال وقتی سن و آدرس به‌عنوان متغیرهای تبیینی برای تمام واحدهای جامعه قبل از نمونه‌گیری موجود باشند ممکن است برای جورسازی استفاده از هر دوی سن و آدرس کار مشکلی باشد، در این صورت ممکن است تنها آدرس (به‌معنای انتخاب همسایه به‌عنوان جانشین) برای جورسازی استفاده شود. به هر حال، هنوز قادریم تفاوت‌های راژمند سن، بین بی‌پاسخ‌ها و جانشین‌ها را توسط مدل‌بندی آماری تعدیل کنیم. در برخی مثال‌ها، متغیرهای تبیینی مدل ممکن است قبل از آمارگیری برای تمام واحدها موجود نباشند اما در طی آمارگیری برای پاسخگویان، بی‌پاسخ‌ها و جانشین‌ها به دست آیند. برای مثال آزمون پرسشنامه اولیه در یک آمارگیری ملی سنجش سلامت و تغذیه (National Health and Nutrition Examination Survey) [۳۶] متغیرهای تبیینی مدل را برای شرکت‌کنندگانی که ارزیابی‌های فیزیکی خود را بازگو نکرده بودند، فراهم کرد. به‌طور مشابه، مرحله اول طرح نمونه‌گیری دو مرحله‌ای می‌تواند متغیرهای تبیینی مدل را فراهم کند.

متغیر تبیینی میدانی متغیری است که می‌تواند در جورسازی استفاده شود ولی نمی‌تواند در مدل‌بندی آماری قرار داده شود. برای مثال اگر جانشین‌ها با جورسازی آدرس انتخاب شوند، بی‌پاسخ‌ها و جانشین‌ها ممکن است در مشخصات دیگری که در همسایگی آن‌ها موجود است مشترک باشند. (برای مثال، مقدار آلودگی صوتی اطراف آن‌ها، یکسان است). این مشخصات ممکن است هم به پاسخ آمارگیری و هم به احتمال پاسخ وابسته باشند. تعدیل‌های بی‌پاسخی سنتی، همچون وزن‌دهی و جانهی، ممکن است نتوانند از این‌گونه متغیرهای کمکی میدانی استفاده کنند زیرا این اطلاعات ثبت نمی‌شوند. به‌علاوه تلاش برای وارد کردن عامل‌هایی همچون آدرس در یک مدل آماری با استفاده از متغیرهای نشانگر برای معرفی رده‌هایی از واحدهای هم‌ارز (مانند همسایگان) نوعی از شیوه جورسازی جانشینی است. در این حالت جانشین‌ها از بین پاسخگویان آمارگیری با

استفاده از جورسازی اندیس رده آن‌ها، و نه آدرس دقیق، انتخاب می‌شوند.

حتی اگر در جامعه مدل (۱) برقرار باشد، نمی‌توان از آن برای جانهی بی‌پاسخی استفاده کرد، زیرا x_f مستقیماً مشاهده نمی‌شود. به هر حال، چون جانشین‌ها از مجموعه واحدهای آماری خارج از نمونه انتخاب می‌شوند، نتیجه می‌شود که رابطه (۱) برای جانشین‌ها نیز برقرار است و در نتیجه جانهی‌ها می‌توانند بر اساس مدل زیر باشند:

$$y_i = y_i^s + \beta_1\left(x_{m,i} - x_{m,i}^s\right) + \varepsilon_i' \qquad (۲)$$

که در آن y_i^s، $x_{m,i}^s$ و $x_{f,i}^s$ به‌ترتیب مقادیر y، x_m و x_f جانشین‌شده برای واحد i را نشان می‌دهند. در این‌جا فرض شده است که در نتیجه فرایند جورسازی $x_{f,i}^s = x_{f,i}$. در عمل این همیشه درست نیست و پژوهش‌های آتی باید رابطه اریبی نتیجه‌شده به علت اندازه خوشه و تمایل به پاسخ‌دهی در داخل خوشه‌ها را به‌طور راژمند ارزیابی کنند. در حالت ساده‌ای که x_f متغیر نشانگری است که عضویت به خوشه را نشان می‌دهد (مانند آن‌چه در شبیه‌سازی روبین و زانوتو [۳۰] آمده است) وقتی اندازه خوشه و احتمال پاسخ عضو خوشه افزایش می‌یابند، باید شانس بدجورسازی در x_f بین جفت بی‌پاسخ‌ها و جانشین‌های آن‌ها کاهش یابد. در برخی آمارگیری‌ها، ممکن است نتوان پاسخگویان به آمارگیری اولیه‌ای را یافت که بر حسب x_f با بی‌پاسخ‌ها جور باشند. در این صورت جانشین‌ها برای بی‌پاسخ‌ها غیر ضروری است. برای برازش مدل (۲) روبین و زانوتو [۳۰] جانشین‌ها را برای برخی از پاسخگویانی که از نظر توزیع‌شان بر حسب x_m با بی‌پاسخ‌ها شبیه بودند، انتخاب کردند (شبیه‌سازی آنان تعداد جانشین‌ها را برابر با $n^* = ۰/۳n_m$ انتخاب می‌کند، که در آن n_m تعداد بی‌پاسخ‌ها است. انتخاب n^* هم بر هزینه آمارگیری و هم بر دقت برآوردگرها تأثیر می‌گذارد.) با فرض وجود رابطه‌ای یکسان بین تفاوت‌های y در (آ) بین پاسخگویان و جانشین‌ها و (ب) بین بی‌پاسخ‌ها و جانشین‌ها، می‌توان داده‌های گم‌شده را بر اساس مدل زیر جانهی چندگانه کرد:

$$y_i \sim N\left(y_i^s + b_0 + b_1(x_{m,i} - x_{m,i}^s), \sigma^2\right) \qquad (۳)$$

که در آن b_0، پارامتر عرض از مبدأ، برای کمک به کاهش بدمشخص‌سازی

(misspecification) مدل انتخاب شده است. برای کاهش وخامت ناشی از نادرستی فرض‌های مدل، جانشین‌ها باید برای آن پاسخگویانی انتخاب شوند که از حیث x_m و x_f شبیه بی‌پاسخ‌ها هستند. این کار بعد فضای متغیرهای کمکی را که در مدل استفاده می‌شوند کاهش می‌دهد. چون استفاده از جورسازی خوب، استفاده از x_f را در مدل غیر ضروری می‌کند، انتخاب پاسخگویانی که از لحاظ توزیع x_m با بی‌پاسخ‌ها مشابه‌اند، بسیار مهم است.

روبین و زانوتو [۳۰] برای روشن‌سازی عملکرد روش خود نسبت به برخی روش‌های رایج مورد استفاده در تعدیل بی‌پاسخی، نتایج چند شبیه‌سازی را خلاصه کرده‌اند. آنان به برآورد میانگین جامعه، $\overline{Y}$، توجه داشتند. خلاصه نتایج شبیه‌سازی روبین و زانوتو [۳۰] آن است که اگر جورسازی امکان‌پذیر باشد و بتوان متغیرهای تبیینی را برای مدل در نظر گرفت از روش MMM وقتی خوشه‌بندی پنهان وجود داشته باشد می‌توان برای تعدیل بی‌پاسخی غیر قابل چشم‌پوشی استفاده کرد. روش MMM در مقایسه با نزدیک‌ترین رقیب این روش یعنی روش پیگیری بی‌پاسخ‌ها، دارای این مزیت است که می‌تواند در حالاتی که امکان پیگیری بی‌پاسخ‌ها وجود ندارد نیز استفاده شود. همچنین در استفاده از اطلاعات انواع گوناگونی از جانشین‌ها انعطاف‌پذیر است. محدودیت کاربردی این روش آن است که باید متغیرهای جورسازی‌ای که با متغیرهای میدانی همبسته‌اند وجود داشته باشند و برخی متغیرهای کمکی که از آن‌ها بتوان برای تفاوت‌های راژمند بین جانشین‌ها و بی‌پاسخ‌ها استفاده کرد، موجود باشند. به‌علاوه هزینه و منطق استفاده از این روش باید در نظر گرفته شوند. برای جزئیات این شبیه‌سازی به روبین و زانوتو [۳۰] یا گنجعلی و دیگران [۲] رجوع شود.

سپاس‌گزاری

از پژوهشکده آمار برای حمایت مالی طرح پژوهشی مرتبط با این مقاله تشکر می‌کنیم. از جناب آقای دکتر حمیدرضا نواب‌پور نیز برای تصحیح و ارائه پیشنهادهای ارزنده در نسخه اولیه این مقاله تشکر می‌کنیم.

مرجع‌ها

[۱] گرامی، عباس؛ پاشا، عین‌اله؛ گنجعلی، مجتبی؛ هرندی، فاطمه؛ رنجی، حسن؛ حمیدی، امید؛ تاپاک، لیلی (۱۳۸۴). گزارش طرح پژوهشی « جانهی بی‌پاسخی‌ها و آثار آن در برآوردها و تجزیه و تحلیل داده‌ها ». پژوهشکده آمار، تهران.

[۲] گنجعلی، مجتبی؛ شفیعی، خلیل؛ رنجی، حسن؛ دانش پرور، نیما؛ رضایی قهرودی، زهرا؛ بیداربخت‌نیا، آرمان (۱۳۸۵). گزارش طرح پژوهشی « بررسی روش‌های کاهش و تعدیل بی‌پاسخی در آمارگیری‌ها ». پژوهشکده آمار، تهران.

[3] Baker, S.G.; Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Amer. Stat. Assoc.* **83**, 62-69.

[4] Biemer, P.; Chapman, D.W.; Alexander, C. (1985). Some research issues in random-digit-dialing sampling and estimation. *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, 71-86, U.S. Bureau of the Census, Washington, D.C.

[5] Birnbaum, Z.W.; Sirken, M.G. (1950). Bias due to non-availability in sampling surveys. *J. Amer. Stat. Assoc.* **45**, 98-111.

[6] Chapman, D.W. (1983). The impact of substitution on survey estimates. In *Incomplete Data in Sample Surveys, vol. II, Theory and Bibliographies*, W.G. Madow, I. Olkin, and D.B. Rubin, eds. Academic Press, New York.

[7] Chapman, D.; Roman, A. (1985). An investigation of substitution for an RDD survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 269-274.

[8] Cochran, W.G. (1977). *Sampling Techniques, 3rd Edition*. Wiley, New York.

[9] Cochran, W.G. (1983). Historical Perspective. In *Incomplete Data in Sample Surveys, vol. II, Theory and Bibliographies*, W.G. Madow, I. Olkin, and D.B. Rubin, eds. Academic Press, New York.

[10] Cohen, R. (1955). *An investigation of modified probability sampling procedures in interview surveys*. Master's thesis, The American University.

[11] Durbin, J.; Stuart, A. (1954). Callbacks and clustering in sample surveys: An experimental study. *J. Roy. Statist. Soc. Ser. A.* **117**, 387-418.

[12] Fay, R.E. (1986). Causal models for patterns of nonresponse. *J. Amer. Stat. Assoc.* **81**, 354-365.

[13] Fowler, F.J. (1993). *Survey Research Methods, 2nd Edition*. Newbury Park, CA: Sage.

[14] Glynn, R.J.; Laird, N.M.; Rubin. D.B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *J. Amer. Stat. Assoc.* **88**, 984-993.

[15] Greenlees, J.S.; Reece, W.S.; Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *J. Amer. Stat. Assoc.* **77**, 251-261.

[16] Groves, R. (1989). *Survey Errors and Survey Costs*. Wiley, New York.

[17] Hansen, M.H. (1975). Comment: Accuracy in market surveys. I: Nonresponse levels and effects. *Proceedings of the Section of Business and Economic Statistics of the American Statistical Association*. 90-92.
[18] Kalton, G.; Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*. **12**, 1-16.
[19] Kish, L. (1965). *Survey Sampling*. Wiley, New York.
[20] Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*. **8**, 183-200.
[21] Kish, L.; Hess, I. (1959). A replacement procedure for reducing the bias of nonresponse. *Amer. Statist.* **13**, 17-19.
[22] Lessler, J.T.; Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. Wiley, New York.
[23] Little, R.J.A.; Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd Edition*. Wiley, New York.
[24] Nathan, G. (1980). Substitution for non-response as a means to control sample size. *Sankhya, Ser. C*. **42**, 50-55.
[25] Nordheim, E.V. (1984) Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. *J. Amer. Stat. Assoc.* **79**, 772-780.
[26] Robins, J.M. (1997). Nonresponse models for the analysis of nonmonotone nonignorable missing data. *Statistics in Medicine*. **16**, 21-37.
[27] Rotnitzky, A.; Robins, J. (1997). Analysis of semi-parametric regression models with nonignorable nonresponse. *Statistics in Medicine*. **16**, 81-102.
[28] Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Amer. Stat. Assoc.* **72**, 538-543.
[29] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
[30] Rubin, D.B.; Zanutto, E. (2002). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations. In *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge, R.J.A. Little, eds. Wiley, New York, pp. 389-402.
[31] Särndal, C.E.; Swensson, B.; Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.
[32] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.
[33] Sirken, M. (1975). Evaluation and critique of household sample surveys of substance abuse. In *Alcohol and other drug use and abuse in the State of Michigan.* Final report, prepared by the Office of Substance Abuse Services, Michigan Department of Public Health.
[34] Sudman, E. (1976). *Applied Sampling*. Academic Press, New York.
[35] Titterington, D.M.; Sedransk, J. (1986). Matching and linear regression adjustment in imputation and observational studies. *Sankhya, Ser. B*. **48**, 347-367.
[36] U.S. Department of Health and Human Services (1994). Plan and operation of

the Third National Health and Nutrition Examination Survey, 1988-94. *Vital Health Statistics, Ser. 1*, No. 32, National Center for Health Statistics, Hyattsville, MD.

[37] Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of Official Statistics.* **15**, 335-350.

[38] Waksberg, J. (1985). Comments: Some research issues in random digit dialing sampling and estimation. *Proceedings of the Bureau of the Census Annual Research Conference.* 87-92, Bureau of the Census, Washington, D.C.

[39] Williams, S.R.; Folsom, R. (1977). Bias resulting from school nonresponse: Methodology and findings. Prepared by the Research Triangle Institute for the National Center for Educational Statistics, Washington DC.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی