مجله علوم آماری، پاییز و زمستان ۱۳۹۷
جلد ۱۲، شماره ۲، ص ۴۶۹ – ۴۸۳
DOI: 10.29252/jss.12.2.469

روش‌های وزن دهی احتمال معکوس و جانهی چندگانه برای تحلیل پاسخ در حالت گمشدگی

فرشته عثمانی، علی‌اکبر راسخی

گروه آمار زیستی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۵/۰۷/۰۲ تاریخ آخرین بازنگری: ۱۳۹۷/۰۴/۲۹

چکیده: ریزش داده‌ها و مقادیر گمشده از مشکلات معمول در تحلیل داده‌ها محسوب می‌شود. لذا اهمیت دارد که با برآورد مقادیر گمشده، داده‌ها کامل شده و در مسیری مناسب و صحیح برای تحلیل قرار داده شوند. دو روش معمول برای مقابله با داده‌های گمشده «جانهی چندگانه» و «وزن‌دهی احتمال معکوس» هستند. در این مقاله، رویکرد سومی معرفی خواهد شد که ترکیبی از دو روش جانهی چندگانه و وزن‌دهی احتمال معکوس است. با توجه به نتایج حاصل از مطالعه شبیه‌سازی روش ترکیبی مزایای بیشتری نسبت به سایر گزینه‌ها دارد. با توجه به وجود مقادیر گمشده در اکثر مطالعات به‌خصوص در حوزه پزشکی، نادیده گرفتن آن‌ها سبب تحلیل اشتباه شده و استفاده از روش‌های نیرومند، برای تحلیل صحیح، با اهمیت تلقی می‌شود.

واژه‌های کلیدی: جانهی چندگانه، وزن‌دهی احتمال معکوس، گمشدگی.

۱ مقدمه

داده‌های گمشده در بسیاری از مطالعات آماری از جمله مدل‌های رگرسیونی وجود دارند و باعث کاهش دقت برآورد می‌شوند. تاکنون روش‌های گوناگونی برای مقابله با مشکل داده‌های گمشده ابداع شده که عموما بر داده‌های گمشده متغیر پاسخ متمرکز بوده است. حال آنکه متغیرهای پیش‌گو نیز می‌توانند دستخوش تغییر و از دست رفتن اطلاعات شوند. انتساب مقادیر مناسب به داده‌های گمشده یکی از چالش‌های موجود در پیش‌پردازش داده‌ها در بسیاری از حوزه‌ها است (یانگ و همکاران ، ۲۰۰۸). طی سال‌های اخیر روش‌های

آدرس الکترونیکی نویسنده مسئول مقاله: فرشته عثمانی، f.osmani@modares.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J12، 62H99

بسیاری برای غلبه بر این مشکلات ارائه شده است. حال آنکه استفاده از این روش‌ها در بسیاری از موارد به دلیل اریبی که وارد مسأله می‌کنند، بیش از آنکه کیفیت داده را افزایش دهد، باعث کاهش کیفیت داده می‌شود.

به‌طورکلی می‌توان چهار راهکار برای برخورد با داده‌های گمشده در نظر گرفت. اولین و ساده‌ترین راهکار، حذف واحدهای دارای مشاهدات ناقص از مطالعه و انجام تحلیل آماری بر اساس اطلاعات واحدهایی است که در تمام زمان‌های اندازه‌گیری در دسترس بوده‌اند. هرچند استفاده از روش تحلیل مبتنی بر موارد کامل (CC) به دلیل سادگی در بین پژوهشگران بسیار رایج است، اما در اکثر مواقع منجر به نتایج اریب و نامعتبر می‌شود. راهکارهای دیگر، روش‌های جانهی، روش‌های وزن‌دهی و روش‌های مبنی بر تابع درستنمایی هستند.

درحالت کلی هر سه روش یاد شده به جانهی مقادیر گمشده با مقادیری خاص می‌پردازند؛ با این تفاوت که جانهی مقادیر گمشده در روش اول به طور واضح و مستقیم، ولی در دو روش دیگر به‌طور غیرمستقیم انجام می‌شود. منظور از جانهی، جایگذاری مقادیری معقول به جای مقادیر گمشده است. بسته به برقراری فرضیات متفاوت در مورد الگوی داده‌های گمشده، روش مورد استفاده نیز متفاوت خواهد بود (شافر و همکاران ، ۲۰۰۳). فعالیت‌های آماری انجام گرفته در معرفی و بررسی مزایا و معایب هر یک از راهکارهای فوق بسیار گسترده است. هم‌چنین کاملا مشخص است که تحلیل های ساده از داده‌های موجود، سبب ایجاد اریبی‌هایی در برآورد پارامترها می‌شود. اما تاثیر دقیق این داده‌های ناکامل بستگی به تعداد مقادیرگمشده و هم‌چنین شدت ارتباط بین متغیرهای پاسخ و کمکی و نشانگرهای داده‌های گمشده دارد (هدکر و گیبنز ، ۲۰۰۶).

عوامل مختلفی می‌توانند به دسترسی پاسخ و متغیرهای کمکی در زمان‌های بررسی تاثیرگذار باشند. معادلات برآورد تعمیم یافته وزنی احتمال معکوس (IPW-GEE) یک روش مناسب برای انجام این کار است. مطالعات تجربی نشان دادند که در نمونه‌های متوسط، اریبی برآوردهای تجربی به‌دست آمده خیلی کوچک هستند (شافر و همکاران ، ۲۰۰۳). روش معادلات برآورد تعمیم یافته (GEE) برای اولین بار توسط لیانگ و زیگر (۱۹۸۶) به عنوان روشی برای تحلیل داده‌های طولی ارائه شد. تحت مکانیسم گمشدگی کاملاً تصادفی (MCAR) تحلیل‌ها براساس GEE برآوردهای سازگاری از پارامترهای رگرسیونی به‌دست می‌دهند. ولی اگر حالت‌های گمشدگی داده‌ها به‌صورت مکانیسم گمشدگی تصادفی (MAR) یا گمشدگی غیرتصادفی (MNAR) باشد، تحلیل‌ها براساس GEE برآوردهای ناسازگاری از پارامترهای رگرسیونی به‌دست می‌دهند (فیتسموریس و همکاران ، ۲۰۰۹).

رابینز و همکاران (۱۹۹۴) گروهی از معادلات برآوردی تعمیم‌یافته وزنی با احتمال معکوس را معرفی کردند که در آن وقتی حالت گمشدگی داده‌ها از نوع مکانیسم تصادفی باشد برآوردهای سازگاری به‌دست می‌دهد. در این روش وزن‌ها از مدل‌های گمشدگی به‌دست آمده‌اند که این مدل‌ها برای دستیابی به برآوردهای سازگار باید به درستی تعیین شوند. از طرفی دو روش رایج در برخورد با داده‌های گمشده، روش‌های MI و IPW هستند. روش IPW برای تعدیل کسرهای نامساوی از نمونه‌گیری قابل استفاده بوده است. به طورکلی روش MI با وجود پیچیدگی کاراتر از IPW است. به‌طوری که روش IPW فقط نیاز به یک مدل احتمال دارد، در حالی که روش MI نیاز به یک مدل توزیع توام از داده‌های گمشده به شرط داده‌های مشاهده شده نیز دارد. مناسب نبودن هر یک از این مدل‌ها در صورتی‌که حجم گمشدگی زیاد باشد، ممکن است منجر به اریبی قابل توجهی شود. در روش IPW مجدد تنها موارد کامل وارد تحلیل می‌شوند و وزن‌ها برای متعادل ساختن مجموعه موارد کامل به منظور نمایندگی از کل نمونه، استفاده می‌شوند.

با توجه به مطالب بیان‌شده، تمرکز مقاله حاضر بر مرور و مقایسه روش جانهی چندگانه و روش وزن‌دهی احتمال معکوس با استفاده از یک مطالعه شبیه‌سازیّریال خواهد بود.

۲ روش IPW-GEE

رابینز و همکاران (۱۹۹۴) تعدیل اریبی ناشی از بی‌پاسخی واحدها را با روش وزن‌دهی معادله‌های برآورد تعمیم‌یافته مولنبرگ انجام دادند که در آن وزن برای هر واحد، به صورت یک ماتریس قطری در نظر گرفته شد. همچنین فیتسموریس و همکاران (۲۰۰۹) نیز نوع دیگری از وزن‌دهی را برای معادله‌های برآورد تعمیم‌یافته لیانگ و زیگر (۱۹۸۶) ارائه کردند که در آن وزن برای هر واحد نمونه‌ای تنها به‌صورت یک عدد در نظر گرفته می‌شد. وقتی گمشدگی MAR باشد، استنباط راجع به میانگین به هر نوع اشتباهی در تعیین توزیع توام بردار پاسخ، بسیار حساس است (دانیلز و هوگان ، ۲۰۰۸). در روش GEE استاندارد لازم است، مدلی برای میانگین مشاهدات به شرط داشتن متغیرهای پیش‌گو وجود داشته باشد. به‌طورکلی، این مدل برای داده‌های مشاهده‌شده از نوع گمشدگی MAR برقرار نیست، بنابراین اعتبار تحلیل‌ها به خطر می‌افتد. با استفاده از روش GEE وزنی ساده، روش‌هایی برای تعدیل تحلیل‌ها ارائه می‌شود، که در آن وزن‌ها با استفاده از یک مدل برای احتمال گمشدگی برآورد می‌شوند و معمولا روش‌های وزنی تمایل نامیده می‌شوند. در این روش وزن W_i را فقط برای افرادی که مطالعه را بدون انصراف به پایان رسانده‌اند، با استفاده از مقادیر پاسخ‌های مشاهده‌شده و متغیرهای کمکی محاسبه می‌کنند. این وزن برای هر فرد

به‌صورت

$$\begin{aligned} W_i &= P(D_i = n+1)^{-1} \\ &= [P(D_i > 1|D_i \geq 1)P(D_i > 2|D_i \geq 2)\dots P(D_i > n|D_i \geq n)]^{-1} \\ &= [\pi_{i1}\pi_{i2}\dots\pi_{in}]^{-1} \end{aligned}$$

محاسبه می‌شود، که در آن

$$\pi_{ij} = P(R_{ij} = 1|R_{ij} = \dots = R_{i,j-1} = 1, X_i, Y_{i1}, \cdots, Y_{i,j-1}).$$

سپس هر فردی که احتمال باقی‌ماندن کمتری داشته باشد (احتمال انصراف بیشتری داشته باشد)، وزن بیشتری را به خود اختصاص می‌دهد. در نهایت روش GEE استاندارد را با این اوزان تطبیق می‌دهند تا براساس یک روش GEE موزون داده‌ها تحلیل شود. پس در این مدل، به افرادی که انصراف نداشته‌اند نیز وزنی برای انصراف داده می‌شود تا براساس این وزن، توزیع شرطی متغیر پاسخ در افرادی که انصراف نداشته‌اند همانند افرادی شود که انصراف داشته‌اند (فیتسموریس و همکاران ، ۲۰۰۴). د رروش GEE استاندارد، از معادلات برآوردگر

$$\sum_{i=1}^{N} D_i' V_i^{-1}(Y_i - \mu_i) = 0$$

استفاده می‌شود، اما در GEE موزون، از معادلات برآوردگر موزون

$$\sum_{i=1}^{N} D_i' V_i^{-1} W_i(Y_i - \mu_i) = 0$$

استفاده می‌شود.

برای برآورد میانگین پاسخ (درحالت گمشدگی در پاسخ) روش IPW نیز قابل استفاده است. این تحلیل وقتی احتمال مشاهده شدن پاسخ فرد به پاسخ مربوطه بستگی نداشته باشد، برآورد سازگاری به‌دست می‌دهد. اما در غیر این صورت اریب خواهد بود (دانیلز و هوگان ، ۲۰۰۸). برآوردگر IPW هورویتز-

تامسون روشی برای تصحیح این اریبی به‌دست می‌دهد، در این روش، مجدد تنها افرادی با مشاهدات پاسخ کامل وارد می‌شوند، اما وزن‌هایی برای متعادل کردن مجموعه موارد کامل استفاده می‌شوند. تا اینکه به عنوان نماینده‌ای از کل نمونه به حساب آیند. وزن هر فرد برابر با معکوس احتمال کامل کردن مطالعه است که این احتمال ناشناخته و نیاز به برآورد دارد (سیمن و همکاران ، ۲۰۱۳).

۳ روش جانهی چندگانه

استفاده از روش‌های جانهی، برای پر کردن جاهای خالی در داده‌هاست؛ به این‌صورت که جاهای خالی حاصل از گمشدگی با مقادیر مناسب جانهی شده پر می‌شوند، یکی از مزایای این رویکرد، این است که با جایگذاری مقادیر گمشده، مجموعه داده‌ها به صورت کامل درآمده و برای تحلیل می‌توان از روش‌های مناسب مربوط به داده‌های کامل استفاده کرد (عثمانی و همکاران ، ۲۰۱۴). در روش جانهی ساده، به‌جای مقادیر گمشده، یک مقدار جانهی شده تک قرار داده می‌شود و در تحلیل با مقدار جانهی شده به همان طریقی رفتار می‌کند که با مقادیری واقعی اندازه‌گیری شده رفتار می‌شود و عدم‌حتمیت این مقادیرگمشده در نظر گرفته نمی‌شود. به همین دلیل، رابینز و وانگ (۲۰۰۰) روشی به نام جانهی‌چندگانه را ابداع کردند، که در آن جای هر مقدار گمشده با دو یا چند مقدار جانهی شده پر شده و هر مجموعه از داده‌های به‌دست آمده با استفاده از روش‌های مربوط به داده‌های کامل تحلیل می‌شود. با ترکیب نتایج این تحلیل‌ها، استنباط‌هایی به‌دست می‌آیند که می‌توانند عدم حتمیت مقادیر گمشده را در نظر بگیرند (روبین ، ۱۹۸۷). در این روش به تعداد جانهی‌های صورت گرفته، مجموعه داده کامل تولید و بر مبنای هر یک از این مجموعه داده‌ها، برآوردی از مولفه‌های مدل و واریانس آن‌هاحاصل می‌شود. سپس در انتها این مقادیر با یکدیگر ترکیب می‌شوند تا برآوردی کلی حاصل شود. از روش‌های رایج جانهی چندگانه می‌توان به روش نمره تمایل، روش مبتنی بر مدل پیش بینی کننده و روش تطابق میانگین پیش‌بینی کننده اشاره کرد. روش جانهی چندگانه دارای مزایا و معایبی است. مزیت اصلی روش جانهی چندگانه همان‌طورکه توسط رابینز و وانگ (۲۰۰۰) اشاره شده، این است که جانهی چندگانه به گردآورندگان داده امکان انعکاس عدم قطعیت مقادیر جانهی شده را می‌دهد. از جمله معایب این روش زمان‌بر بودن جانهی چندین مجموعه داده کامل، آزمون جداگانه هر کدام از این مجموعه داده‌ها و همچنین ترکیب نتایج این مدل‌ها است. در روش MI، داده‌های گمشده با داده‌های به‌دست آمده از مدل جانهی جایگزین می‌شوند. این‌عمل M بارتکرار می‌شود، درنتیجه M مجموعه داده کامل بوجود می‌آید. هرکدام از این مجموعه‌ها، جداگانه تحلیل می‌شوند و برآوردی از پارامترهای مدل θ به‌دست می‌دهند. فرض کنیم $\hat{\theta}$ برآوردگر θ از مجموعه داده کامل، $\hat{V}$ واریانس برآورد شده و $\hat{\theta}_{(m)}$ و $\hat{V}_{(m)}$

مقادیر برآورد شده از m امین مجموعه داده جانهی شده باشند ($m = ۱, \ldots, M$). برآورد و واریانس آن به ترتیب عبارتند از

$$\hat{\theta}_M = \frac{۱}{M}\sum_{i=۱}^{M}\hat{\theta}_M,$$
$$\hat{V}_M = \frac{۱}{M}\sum_{i=۱}^{M}\hat{V}_{(M)} + (۱ + M^{-۱})(M-۱)^{-۱}\sum_{i=۱}^{M}(\hat{\theta}_{(m)} - \hat{\theta}_M)(\hat{\theta}_{(m)} - \hat{\theta}_M)^T.$$

برای به‌دست آوردن برآورد پارامترها در این روش از قوانین ادغام روبین (۱۹۸۷) استفاده شده و برآورد واحدی برای پارامتر رگرسیونی با معدل گرفتن از M برآورد حاصله به‌دست می‌آید. واریانس براورد نیز با ترکیب واریانس درون جانهی و واریانس بین جانهی به‌دست می‌آید. یعنی

$$\hat{\beta} = \frac{۱}{M}\sum_{k=۱}^{m}\hat{\beta}^{(k)} = \bar{\beta}$$

و

$$\frac{۱}{M}\sum_{k=۱}^{m}\hat{\mathrm{Cov}}(\hat{\beta}^k) = (۱ + \frac{۱}{m})\frac{۱}{m-۱}\sum_{k=۱}^{m}(\hat{\beta}^k - \bar{\beta})(\hat{\beta}^k - \bar{\beta})^T.$$

روش‌های IPW و MI درحالت گمشدگی تصادفی و در صورتی‌که مدل وزندهی و جانهی به‌درستی مشخص شده باشد، برآوردگرهای سازگاری از θ به‌دست می‌دهند. در روش MI وقتی‌که $\hat{\theta}$ برآوردگر MLE، ماتریس $\hat{V}$ معکوس ماتریس اطلاع فیشر و همچنین داده‌های گمشده از توزیع پسین بیزی نمونه‌گیری شده باشند، استنباط می‌شود که $\hat{\theta}_M$ به طور مجانبی دارای توزیع نرمال با واریانس V و $\hat{V}_M$ برآوردگر نااریب V_M است و در صورتی که $M = \infty$، این برآوردگر سازگار نیز خواهد بود (فیتسموریس و همکاران ، ۲۰۰۴).

اگر مدل جانهی به‌درستی مشخص شده باشد، روش MI اغلب به روش IPW به علت کارایی بیشتر ترجیح داده می‌شود. ولی در صورتی که مقادیرگمشده جانهی شده زیاد باشند، ممکن است منجر به اریبی قابل توجهی شود؛ اما اگر متغیرهای کمی دارای مقادیر گمشده باشند، جانهی این مقادیر نسبت به خارج کردن آن‌ها رضایت‌بخش‌تر خواهد بود. از طرف دیگر، اگر متغیرهای زیادی از افراد دارای مقادیر گمشده باشند؛ پیچیدگی مدل و احتمال انتخاب نادرست مدل جانهی ممکن است نگران‌کننده باشد. این موقعیت ممکن است به‌عنوان مثال درمطالعات طولی، زمانی‌که کلیه بلوک‌های داده‌ها در بعضی از افراد به‌دلیل از

دست دادن ملاقات‌ها گمشده باشند یا در مطالعه‌ای که بعضی از افراد همه سوال‌های مربوطه را پاسخ نمی‌دهند، پیش آید که در چنین موقعیت‌هایی تحلیل‌گر اعتماد بیشتری به استفاده از روش IPW دارد (شافر ، ۱۹۹۷).

۴ مطالعه شبیه‌سازی

در این بخش، عملکرد تجربی روش‌های ارائه شده از طریق یک مطالعه شبیه‌سازی بررسی و با سایر روش‌ها که ممکن است در عمل استفاده شوند، مقایسه می‌شود. .همچنین از روش IPW/MI برای جانهی پاسخ در یک مدل رگرسیونی استفاده خواهد شد.

۱.۴ مدل شبیه‌سازی

مدل تحلیل، تنها برای افرادی با متغیرهای کمکی کامل، برازش داده شده و متغیرپاسخ در این افراد جانهی خواهد شد. برای تحلیل نمونه تولید شده دو مرحله گمشدگی در X و Y وجود دارد. در مرحله اول یا می‌توان افراد با X ناکامل را حذف کرد یا اینکه مقادیر گمشده X را جانهی نمود. به طور مشابه، هر فردی با پاسخ گمشده که در مرحله اول حذف نشده، می‌تواند در مرحله دوم حذف یا جانهی شود. قابل ذکر است که در هر مرحله اگر حذف انجام شود می‌توان از روش IPW برای تعدیل داده‌ها استفاده کرد. بنابراین در هر مرحله سه حالت و در کل $۳ \times ۳ = ۹$ حالت وجود دارد. اگر راهکارهای بکار گرفته شده در این دو مرحله با نماد ST1/ST2 نشان داده شود، این راهکارها می‌توانند یکی از روش‌های مبتنی بر موارد کامل (حذف بدون وزن دهی)، IPW (حذف با وزن‌دهی بقیه) یا MI باشند.

در روش IPW/MI که هدف اصلی این مقاله است افراد با متغیرهای گمشده X حذف می‌شوند و تعدیل داده‌ها با وزن‌دهی سایر داده‌ها صورت می‌گیرد. همچنین درافراد با X کامل اما با گمشدگی در متغیر پاسخ، Y جانهی می‌شود و در روش CC/CC هم تنها افرادی که X و Y کامل دارند بدون وزن‌دهی تحلیل می‌شوند و در روش IPW/IPW هم همین افراد استفاده می‌شوند، با این تفاوت که در این حالت، افراد به اندازه معکوس احتمال باقی ماندن در مطالعه وزن‌دهی می‌شوند، همچنین در روش MI/MI همه مقادیر گمشده جانهی می‌شوند. به‌علاوه روش‌های CC/IPW، CC/MI و IPW/CC هم در نظر گرفته شدند.

۲.۴ تولید داده‌ها

انجام این شبیه‌سازی با سه هدف دنبال شد: الف- بررسی نااریب بودن $\hat{V}$ برای روش IPW/MI. ب- بررسی کارایی روش IPW/MI نسبت به روش IPW/IPW. ج- نشان دادن اینکه روش MI/MI زمانی‌که مدل جانهی به‌درستی تعیین نشود، برآوردگرهای اریبی را به‌دست می‌دهد و IPW/MI به‌طور مجانبی نااریب و یا حداقل از روش MI/MI دارای اریبی کمتری است. با توجه به این اهداف فرایند تولید داده‌ها به صورت زیر در نظر گرفته شد.

بردار $X = (x_۱, \ldots, x_۵)$ و متغیر Y برای $M = ۱۰۰۰$ فرد تولید شد. برای هر فرد، $x_۱$ با احتمال ۰٫۵ مقدار یک و در غیر این صورت صفر را اختیار می‌کند. مشاهدات $(x_۲, x_۳, x_۴)$ به‌صورت مستقل از توزیع $N(۰, ۱)$ تولید شدند. همچنین $x_۵$ از توزیع $N(x_۲x_۳, ۱)$ تولید شد. متغیر پاسخ نیز از رابطه

$$Y = -۳ + x_۱x_۲ + x_۱x_۳ + ۰/۵x_۲x_۳ + x_۴ + ۰/۵x_۵ + \epsilon, \qquad \epsilon \sim N(۰, ۱) \qquad (۱)$$

محاسبه شد. از بین متغیرها $x_۱$ برای همه N نفر مشاهده شده بود ولی با $X = (x_۲, x_۳, x_۴, x_۵)$ احتمال $۰/۸ - ۰/۶x_۱$ مشاهده شدند. از طرفی اگر بردار $X = (x_۲, x_۳, x_۴, x_۵)$ مشاهده می‌شد متغیر پاسخ Y نیز با احتمال $۱ + \exp(-۱٫۵ + ۰٫۶x_۲x_۴)$ مشاهده و در غیر این صورت گمشده در نظر گرفته شد. مدل تحلیل نیز

$$Y = \theta_۰ + \theta_۲x_۲ + \theta_۳x_۳ + \theta_{۲۳}x_۲x_۳ + e$$

بود، که در آن $E(e|x_۲, x_۳) = ۰$. بدین ترتیب بردار Z به صورت $Z = (۱, x_۲, x_۳, x_۲x_۳)$ در نظر گرفته شد که با انتگرال‌گیری از رابطه (؟؟) نسبت به $x_۱, x_۴, x_۵$ مدل تحلیل به‌درستی مشخص خواهد شد. مقادیر θ هم به‌صورت $(\theta_۰, \theta_۲, \theta_۳, \theta_{۲۳}) = (-۳, ۰٫۵, ۰٫۵, ۱)$ در نظر گرفته شد. این فرایند تولید داده به سه دلیل انتخاب شده است:

۱- اثرات متقابل $x_۱x_۲, x_۱x_۳$ در رابطه (؟؟) نشان دهنده ارتباط بین Y و $x_۲x_۳$ در دو طبقه $x_۱$ است. هم‌چنین احتمال مشاهده در یک طبقه $x_۱$ برابر ۰٫۲ و در طبقه دیگر برابر ۰٫۸ است. بنابراین ارتباط بین Y و $x_۲x_۳$ در افراد با X های کامل و افراد با X های ناکامل با یکدیگر فرق می‌کند. از طرفی تعدیل نکردن گمشدگی در مرحله اول باعث اریبی $\theta_۲$ و $\theta_۳$ می‌شود. بنابراین روش‌های CC/IPW، CC/MI و CC/CC اریب خواهند بود.

۲- برای افرادی با بردار مشاهده شده $(x_۲, x_۳, x_۴, x_۵)$ ، احتمال مشاهده Y به $x_۴$ بستگی دارد که در مدل تحلیل نیست ولی با Y در ارتباط است؛ این مسئله باعث می‌شود که ارتباط بین Y و X در مدل تحلیل، در مجموعه داده کامل با مجموعه داده با X کامل ولی با گمشدگی در Y متفاوت باشد. بخصوص به دلیل اینکه احتمال گمشدگی در Y به $x_۲x_۴$ بستگی دارد. ارتباط بین Y و $x_۲$ در دو مجموعه متفاوت خواهد بود. از طرفی تعدیل نشدن گمشدگی در مرحله ۲ منجر به اریبی بخصوص در $\theta_۲$ خواهد شد.

۳- مقدار $x_۵$ که در فرایند تولید داده برای Y بکار رفته بود، نشان می‌دهد که استفاده از روش MI در مرحله یک در صورتی که مدل جانهی X به‌درستی تعیین نشده باشد می‌تواند باعث اریبی شود (نتایج MI/MI در جدول ۱). یک مجموعه ۱۰۰۰ تایی داده تولید شده و این هفت روش برای هر یک از مجموعه داده‌ها بکار گرفته شد. برای هریک از $(\theta_۰, \theta_۲, \theta_۳, \theta_{۲۳})$ و هر یک از روش‌ها، میانگین و واریانس ۱۰۰۰ پارامترهای برآورد شده محاسبه شدند. مقدار SE نیز به صورت انحراف معیار برآورد پارامترها محاسبه شد. هم‌چنین در روش هایی که شامل روش جانهی بودند ده بار جانهی انجام شد.

۳.۴ تحلیل داده‌ها با روش ترکیبی جانهی چندگانه

در روش MI/MI مدل جانهی در مرحله ۱ به صورت

$$(X_۲, X_۳, X_۴) \sim N\big((\gamma_۲, \gamma_۳, \gamma_۴), \Sigma_۱\big)$$
$$X_۵|X_۲, X_۳ \sim N(\gamma_۵ + \gamma_۶X_۲ + \gamma_۷X_۳ + \gamma_۸X_۲X_۳, \Sigma_۲)$$

در نظر گرفته شد. از پیشین‌های ویشارت معکوس و نرمال استفاده شد که باعث می‌شود پسین‌ها هم نرمال و ویشارت معکوس به‌دست بیاید. برای روش‌های CC/MI و IPW/MI مدل جانهی استفاده شده در مرحله ۲ بصورت زیر است.

$$\begin{aligned} Y &= \beta_۰ + \beta_۱X_۱ + \beta_۲X_۲ + \beta_۳X_۳ + \beta_۴X_۴ + \beta_۵X_۵ \\ &+ \beta_{۱۲}X_۱X_۲ + \beta_{۱۳}X_۱X_۳ + \beta_{۲۳}X_۲X_۳ + \beta_{۱۲۳}X_۱X_۲X_۳ + \epsilon \end{aligned}$$

۴.۴ تحلیل داده‌ها با روش ترکیبی وزن‌دهی احتمال معکوس

برای روش های IPW/MI، IPW/CC و IPW/IPW وزن‌هایی با برازش مدل گمشدگی برای مرحله ۱ برآورد شدند، توجه شود که چون X_1 دو حالتی است تابع

$$W = (\delta_0 + \delta_1 X_1)^{-1} = \delta_0^{-1} - \delta_1 X_1 [\delta_0 (\delta_0 + \delta_1)]^{-1}$$

یک تابع خطی از X_1 است. از این رو مدل جانهی مرحله ۲ شامل $X_1 Z$ و Z است که شامل WZ است. برای روش‌های CC/IPW و IPW/IPW وزن‌هایی با استفاده از مدل صحیح مشخص شده در مرحله ۲ به‌صورت

$$\text{logit}\, P(Y \text{ observed} \,|X \text{ observed}) = \delta_2 + \delta_3 X_2 + \delta_4 X_4 + \delta_5 X_2 X_5$$

براورد می‌شود. در روش IPW/IPW احتمال کامل بودن یک فرد، برابر ضرب این دو احتمال IPW است. برآورد میانگین پارامترها و مقادیر انحراف معیار و هم‌چنین جذر میانگین واریانس‌های برآورد شده در جدول ۱ نشان داده شده است. با توجه به نتایج به‌دست آمده می‌توان گفت بطور مجانبی برآوردگرهای نااریبی از پارامترها و انحراف معیارها به‌دست آمده است.

۵.۴ مقایسه روش‌های ترکیبی

همان‌طور که در بخش پیش بیان شد، روش های CC/IPW، CC/MI و IPW/CC برای حداقل یکی از پارامترها اریب هستند. ولی می‌توان گفت روش‌های IPW/IPW و MI/MI تقریبا نااریب هستند. مدل MI/MI کارایی کمتری نسبت به روش IPW/MI دارد زیرا مدل جانهی در راهکار دوم از اطلاعات کمکی نظیر متغیرهای کمکی به‌ویژه (X_4, X_5) که درمدل تحلیل وارد نشدند، استفاده می‌کند. کاراترین روش نااریب با توجه به نتایج به‌دست آمده روش MI/MI است. جانهی، زمانی که مدل جانهی به‌درستی مشخص شده باشد بهترین روش است ولی زمانی که مدل جانهی در مرحله ۱ یا ۲ به‌درستی مشخص نشود، روش MI/MI ممکن است اریب باشد. همان طور که نیمی از افراد دارای X های ناکامل بودند با برازش مدل تحلیل به کل نمونه برآورد θ_{23} در حدود ۰/۷۵ به‌دست آمد. این موضوع در سطر MI*/MI در جدول ۱ نشان داده شده است (نماد MI* مدل جانهی نادرست مشخص شده را نشان می‌دهد). از طرفی با فرض

اینکه مدل جانهی در مرحله ۱ به‌درستی مشخص شده باشد ولی در مرحله ۲ به دلیل حذف $\beta_{۲۳}X_۲X_۳$، $\beta_{۱۲۳}X_۱X_۲X_۳$ و $\beta_۵X_۵$ به‌درستی مشخص نشده باشد، مقادیرگمشده Y با این فرض که هیچ اثر متقابلی بین $X_۲$ و $X_۳$ نیست، جانهی خواهد شد. حدود ۶۰ درصد از مقادیر Y گمشده هستند و به همین دلیل مقادیر $\theta_{۲۳}$ کم برآورد خواهند شد. این نتایج در ردیف MI/MI* جدول ۱ نشان داده شده است. ردیف IPW/MI* نتایج مربوط به روش IPW/MI با مدل جانهی به اشتباه مشخص شده در مرحله ۲ رانشان می‌دهد. این روش به طور قابل توجهی نسبت به روش MI/MI* اریبی کمتری خواهد داشت، زیرا مقادیر Y کمتری جانهی شده اند. بنابراین قسمت IPW در روش IPW/MI یک روش محافظه کار در برابر انتخاب نادرست مدل جانهی می‌باشد.

بسیاری از نرم افزارهای آماری، دارای بسته‌ها و روش‌هایی برای تحلیل مقادیر گمشده هستند که در این مقاله از نرم افزار R نسخه ۳/۱/۰ و نرم افزار SAS استفاده شد.

بحث و نتیجه‌گیری

مشکل وجود داده‌های گمشده بخصوص در مطالعات پزشکی، در سالهای اخیر به شدت مورد توجه قرار گرفته است (عثمانی و همکاران ، ۲۰۱۵). این مقاله با مقایسه رویکردهای مختلف به دنبال یافتن بهترین رویکرد در مواجهه باگمشدگی است. در تحقیق صورت گرفته، هدف مقایسه ترکیبی روش‌های معادله برآورد وزنی معکوس و روش جانهی چندگانه بود. به بیان دیگر، در این مقاله از ترکیب رویکردهای داده‌های کامل، جانهی چندگانه و وزن‌دهی احتمال معکوس با هدف مقایسه این روش‌هادر تحلیل داده‌ها استفاده شد. جدول ۱ برازش مدل رگرسیونی مفروض را بر داده‌های تولید شده با استفاده از رویکردهای یاد شده نشان می‌دهد. همان‌طور که مشاهده می‌شود ترکیب رویکردهای مختلف، نتایجی متفاوت را از لحاظ اریبی برآورد پارامترها نشان می‌دهند.

بطور کلی، رویکرد استفاده از داده‌های کامل، روشی مناسب برای حل مشکل گمشدگی در داده‌ها نیست، زیرا باعث کاهش شدید حجم نمونه و در نتیجه، کارایی برآوردها می‌شود (مولنبرگ و همکاران ، ۲۰۰۸). رابینز و وانگ (۲۰۰۰) فرمولی کلی برای واریانس برآوردگر MI براساس برآوردگر داده‌های کامل با حل مجموعه‌ای از معادلات برآوردگر ارائه کردند. این فرمول زمانی که جانهی نامناسب و مدل جانهی پارامتری باشد، استفاده می‌شود. درروش IPW/MI نیز این رابطه می‌تواند استفاده شود.

جدول ۱: میانگین برآورد پارامتر، جذر میانگین واریانس برآورد شده (aSE) و براورد تجربی (eSE) برای چهار پارامتر و ده روش تحلیل

$\theta_{۲۳}=۱$			$\theta_۳=-۰/۵$			$\theta_۲=-۰/۵$			$\theta_۰=-۳$			
eSE	aSE	میانگین	eSE	aSE	میانگین	eSE	aSE	میانگین	eSE	aSE	میانگین	روش
۰/۰۹۱	۰/۰۸۲	۱/۰۰۵	۰/۰۸۷	۰/۰۸۰	۰/۲۰۰	۰/۰۸۷	۰/۰۷۸	۰/۰۸۹	۰/۰۷۹	۰/۰۸	−۲/۸۹۵	CC/CC
۰/۱۰۷	۰/۰۹۶	۱/۰۰۴	۰/۰۸۹	۰/۰۸۶	۰/۲۰۰	۰/۰۹۱	۰/۰۸۹	۰/۱۹۶	۰/۰۷۹	۰/۰۸۲	−۲/۸۹۳	CC/IPW
۰/۰۸۶	۰/۰۸۴	۱/۰۰۴	۰/۰۸۳	۰/۰۷۹	۰/۲۰۳	۰/۰۸۳	۰/۰۸۱	۰/۲۰۱	۰/۰۷۵	۰/۰۷۵	−۲/۸۹۴	CC/MI
۰/۱۱۹	۰/۱۱۳	۱/۰۰۷	۰/۱۱۴	۰/۱۰۹	۰/۴۹۵	۰/۱۱۱	۰/۱۱۰	۰/۳۷۸	۰/۱۰۴	۰/۱۰۲	−۲/۸۹۳	IPW/CC
۰/۱۳۲	۰/۱۲۱	۱/۰۰۶	۰/۱۱۶	۰/۱۱۲	۰/۴۹۳	۰/۱۲۲	۰/۱۱۸	۰/۴۸۶	۰/۱۰۸	۰/۱۰۶	−۲/۸۹۱	IPW/IPW
۰/۱۱۲	۰/۱۰۹	۱/۰۰۸	۰/۱۰۷	۰/۱۰۴	۰/۴۹۸	۰/۱۰۳	۰/۱۰۳	۰/۴۹۳	۰/۰۹۴	۰/۰۹۷	−۲/۸۹۲	IPW/MI
۰/۰۸۲	۰/۰۹۲	۱/۰۰۶	۰/۰۸۸	۰/۰۹۰	۰/۴۹۷	۰/۰۸۷	۰/۰۹۲	۰/۴۹۹	۰/۰۷۹	۰/۰۸۹	−۲/۸۹۶	MI/MI
۰/۰۸۴	۰/۱۰۲	۰/۷۴۹	۰/۰۹۴	۰/۰۹۴	۰/۴۹۶	۰/۰۹۳	۰/۰۹۵	۰/۴۹۶	۰/۰۸۳	۰/۹۲	−۲/۹۹۱	MI*/MI
۰/۰۵۵	۰/۰۹۱	۰/۳۹۲	۰/۰۵۱	۰/۰۸۸	۰/۰۹۹	۰/۰۵۲	۰/۰۸۸	۰/۱۰۰	۰/۱۰۲	۰/۱۰۷	−۲/۹۹۴	MI/MI*
۰/۱۲۷	۰/۱۳۲	۰/۷۷۶	۰/۱۱۵	۰/۱۱۷	۰/۴۹۵	۰/۱۲۱	۰/۱۱۹	۰/۴۹۱	۰/۱۰۱	۰/۱۰۶	−۲/۹۹۳	IPW/MI*

در مورد رگرسیون خطی با اعمال روش MI برای پاسخ گمشده، برآوردگر واریانس روبین برای روش IPW/MI وقتی M به سمت بینهایت میل کند، سازگار است. شافر و همکاران (۲۰۰۳) نیز اشاره به این نکته کرده است که اگرچه اثبات عملکرد مناسب روش MI مشکل است، ولی برخی از پژوهشگران ممکن است استفاده از روش MI/MI را به شرطی که مدل‌های جانهی به‌درستی مشخص شده باشند، ترجیح بدهند. در نتیجه این روش کارآمدتر از روش IPW/MI خواهد بود (شافر و همکاران ، ۲۰۰۳). با این‌حال، اگر نتایج حاصل از IPW/MI و MI/MI بسیار متفاوت باشند، تحقیق بیشتر در راستای تصحیح مدل جانهی لازم خواهدبود، زیرا تشخیص نادرست مدل جانهی در MI/MI معمولا باعث اریبی می‌شود.

علاوه براین، روش IPW/MI وقتی مدل وزن‌دهی در مقایسه با مدل جانهی نسبتا ساده‌تر باشد، مطلوبتر خواهد بود. روش جایگزین دیگری برای IPW/MI روش IPW/IPW است که با وجود ساده‌تر بودن، دارای این نقطه ضعف است که اگر یک فرد حتی فقط یک متغیر از او گمشده باشد از مطالعه و تحلیل حذف می‌شود (رابینز و گیل ، ۱۹۹۷). از سوی دیگر IPW/MI امکان استفاده از یک مجموعه از وزن‌ها را می‌دهد.

یافته‌های این مطالعه نشان داد که ترکیب IPW/MI برای تحلیل داده‌های گمشده در صورتی که مدل وزن‌دهی به نسبت مدل جانهی ساده‌تر باشد، بهتر خواهد بود. از طرفی با توجه به اینکه مقادیر گمشده همواره مشاهده نشده باقی خواهند ماند، تمام روش‌های موجود برای تحلیل داده‌های گمشده شامل فرضیات غیرقابل شناسایی و غیرقابل اثبات هستند، بنابراین بهتر است در برخورد با این داده‌ها صرفاً به نتایج حاصل از یک روش اتکا نشود و در این راستا تحلیل حساسیت مناسب نیز انجام شود.

تقدیر و تشکر

نویسندگان از پیشنهادات ارزنده داوران گرامی و ویراستار مجله که باعث ارائه بهتر و بهبود مقاله شده است کمال تشکر و قدردانی را اعلام می‌دارند.

مراجع

Daniels, M. J. and Hogan, J. W. (2008), *Missing Data in Longitudinal Studies*, London, Chapman & Hall.

Diggle, P., Heagerty, P., Liang K. Y. and Zeger. S. (2002), *Analysis of Longitudinal Data*, Second Edition, Oxford University Press.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004), *Applied Longitudinal Analysis*, Wiley-Interscience.

Fitzmaurice, G. M., Davidian M., Verbeke, G. and Molenberghs, G. (2009), *Longitudinal Data Analysis*, Chapman Hall/CRC.

Hedeker, D. and Gibbons R. D. (2006), *Longitudinal Data Analysis*, New York, Wiley.

Liang, K. and Zeger, S. L. (1986), Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, **73**, 13-22.

Lipsitz S. R., Fitzmaurice J. G., Ibrahim M., Sinha D. M. Parzen, M. and Lipshultz, S. (2009), Joint Generalized Estimating Equations for Multivariate Longitudinal Binary Outcomes with Missing Data: An Application to Acquired Immune Deficiency Syndrome Data. *Journal of the Royal Statistical Society*, Series A, **1**, 3-20.

Molenberghs, G., Kenward, M. G. and Goetghebeur, E. (2001), Sensitivity Analysis for Incomeplete Contingency Tables: The Slovenian Plebiscite Case. *Journal of Applied Statistics*, **50**, 15-29.

Molenberghs, G., Beunckens, C., Sotto, C. and Kenward, M. G. (2008), Every Missingness Not at Random Model Has a Missingness at Random Counterpart with Equal Fit, *Journal of the Royal Statistical Society*, Series B, **70**, 371-88.

Robins, J., Rotnitzky, A. and Zhao, L. (1994), Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, **89**, 846-866.

Robins, J. M. and Gill, R. D.(1997), Non-response Models for the Analysis of Non-monotone Ignorable Missing Data. *Statistics in Medicine*, **16**, 39–56.

Osmani, F., Hajizadeh, E. and Mansoori, P. (2014), Analysis of Risk Factors for Psoriasis Recurrence Using Proportional Rates Model, *Journal of Skin and Stem Cell*, **3**, 129-137.

Osmani, F., Hajizadeh, E. and Mansoori, P. (2015), Estimation of Seasonal Effect on the Psoriasis Recurrence Using Time Dependent Coefficient Rates Model for Recurrent Events, *Dermatology and Cosmetic*, **6**, 23-30.

Robins, J. and Wang, N. (2000), Inference for Imputation Estimators, *Biometrika*, **87**, 113–-24.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Seaman, S. R., White, I. R., Copas, A. J. and Li, L. (2012), Combining multiple imputation and inverse-probability weighting, *Biometrics*, **68**, 129-137.

Seaman, S. R. and White, I. R. (2013), Review of Inverse Probability Weighting for Dealing with Missing Data, *Statistical Methods in Medical Research*, 278-295.

Schafer J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.

Schafer, J. L. (2003), Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Statistica Neerlandica*, **57**, 19-35.

Vansteelandt, S., Carpenter, J. and Kenward, M. G, (2010), Analysis of Incomplete Data Using Inverse Probability Weighting and Doubly Robust Estimators, *Methodology* **6**, 37-48.

Yang X., Li J. and Shoptaw S. (2008), Imputation-based Strategies for Clinical Trial Longitudinal Data with Nonignorable Missing Values. *Statistics in Medicine*, **27**, 2826-2849.