مجله تخصصی اپیدمیولوژی ایران؛ ۱۳۹٤، دوره ۱۱، شماره ۳: صفحات ۵۱-٤۳.

مقاله پژوهشی

مقایسه روش الگوریتم EM و روش‌های متداول جانهی داده‌های گمشده: مطالعه‌روی پرسشنامه خوددرمانی بیماران دیابتی

علیرضا افشاری‌صفوی[۱]، حسین کاظم‌زاده قره‌چبق[۲]، منصور رضایی[۳]

[۱] گروه آمار و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی اصفهان

[۲] اداره کل آمار، اداره کل شرق تهران بزرگ ، سازمان تأمین اجتماعی

[۳] گروه آمار و اپیدمیولوژی، دانشکده بهداشت، مرکز تحقیقات توسعه اجتماعی و ارتقاء سلامت دانشگاه علوم پزشکی کرمانشاه

نویسنده رابط: حسین کاظم‌زاده قره‌چبق، نشانی: تهران، خیابان شهید مطهری، انتهای خیابان نور، خیابان هفتم، اداره‌کل تامین اجتماعی شرق تهران بزرگ، اداره‌کل آمار، تلفن: ۸۸۷۴۸۹۳۳،

پست الکترونیک: kazemzadeh_hk@yahoo.com

تاریخ دریافت: ۹۳/۰۷/۳۰؛ پذیرش: ۹۴/۰۶/۰۷

مقدمه و اهداف: داده‌های گمشده، چالش بزرگی در پژوهش‌ها به‌شمار می‌آیند. به فراخور نوع مطالعه و نوع متغیرهای مورد بررسی، روش‌های گوناگونی برای کار با این داده‌ها تا کنون معرفی شده است. هدف این مطالعه مقایسه پنج روش جانهی متداول در برخورد با گمشدگی در داده‌های پرسشنامه‌ای بود.

روش کار: در این مطالعه تعداد ۵۰۰ پرسشنامه مربوط به خوددرمانی در بیماران دیابتی مورد استفاده قرار گرفت. گمشدگی در مشاهده‌ها به‌صورت تصنعی و با انتخاب تصادفی سؤالاتسؤالات و سپس حذف آن‌ها تولید شد. پنج روش جانهی عبارت بودند از: ۱- میانگین سؤالاتسؤالات؛ ۲- میانگین فردی؛ ۳- نمای فردی؛ ۴- رگرسیون خطی؛ و ۵- الگوریتم EM. برای هر روش میانگین و انحراف معیار نمرات جانهی شده با مقادیر اصلی مقایسه گردید. هم‌چنین ضریب همبستگی اسپیرمن، درصد دسته‌بندی اشتباه و آماره کاپا نیز محاسبه شد.

یافته ها: مقدار آماره کاپای بالاتر از ۰/۸۱ برای سطح گمشدگی ۱۰ درصد بیانگر توافق تقریباً کامل در این سطح از گمشدگی بود. الگوریتم EM بالاترین میزان توافق با نتایج داده‌های واقعی را با مقدار آماره کاپای ۰/۸۸۶ نشان داد. هم‌چنین با افزایش میزان گمشدگی اطلاعات به ۳۰ درصد، الگوریتم EM و روش میانگین فردی با مقدار کاپای ۰/۶۹۷ و ۰/۶۸۷ از توافق نسبتاً مشابهی برخوردار بودند.

نتیجه‌گیری: در این مطالعه الگوریتم EM دقیق‌ترین روش برای کار با داده‌های گمشده در تمام الگوهای مورد ارزیابی شناخته شد. روش میانگین فردی به دلیل سادگی کار با داده‌های گمشده به‌ویژه برای بیش‌تر خوانندگان غیرآماری می‌تواند مورد توجه قرار گیرد.

واژگان کلیدی: الگوریتم EM، داده‌های گمشده، دیابت، خوددرمانی، آماره کاپا، رگرسیون

مقدمه

داده‌های گمشده یکی از چالش‌های رایج در پژوهش‌های علوم پزشکی به شمار می‌آید و این مسأله به‌ویژه در مطالعه‌هایی که از ابزارهای گزارش فردی هم‌چون پرسشنامه استفاده می‌کند، متداول‌تر است (۳-۱). پژوهشگری که با این مسأله روبه‌رو می‌شود؛ دو انتخاب دارد، حذف مشاهده‌های دارای گمشدگی و یا استفاده از روش‌های جانهی. در سال‌های اخیر روش‌های مختلفی به منظور جانهی داده‌های گمشده معرفی شده است (۶-۲،۴). هر یک از این روش‌ها به فراخور نوع مطالعه و نوع متغیرها مورد بررسی کارایی خاص خود را دارد. بی‌پاسخی در پرسشنامه‌ها یکی از مشکلات عمده‌ی استفاده از این ابزار جمع‌آوری اطلاعات به شمار می‌آید. عدم دقت کافی در پاسخ به سؤالات، بی‌دقتی در ورود اطلاعات از پرسشنامه به محیط نرم‌افزار، سردرگمی پاسخ‌دهنده در قبال پاسخ به برخی از سؤالات مبهم و ... از شایع‌ترین علل بی‌پاسخی و گمشدگی داده‌های پرسشنامه‌ای به شمار می‌آید. از طرفی اگر این گمشدگی‌ها درصد قابل توجهی از داده‌ها را به خود اختصاص دهد، می‌توان بر نتایج پژوهش تأثیر گذارد. بسیاری از پژوهشگران به دنبال روش‌هایی برای غلبه بر این مشکل می‌باشند. روش‌هایی که ضمن سادگی استفاده و فهم راحت، نتایج قابل اعتمادی را نیز به دنبال داشته باشد.

در این مطالعه از پرسشنامه خوددرمانی بیماران دیابتی به منظور مقایسه پنج روش متداول جانهی داده‌های گمشده استفاده شد که به نوعی تقابل بین سادگی روش و عملکرد روش را به چالش می‌کشد. چهار روش اول (میانگین سؤالات، میانگین فردی، نمای فردی و رگرسیون)، از روش‌های کلاسیک برای مقابله با

گمشدگی به شمار می‌آیند، که از نظر تئوری برای افراد غیر آماری نیز قابل درک می‌باشند. روش آخر (الگوریتم EM) یکی از روش‌های مدرن و پیشرفته در حل مسأله گمشدگی تلقی می‌شود که از دیدگاه تئوری از پیچیدگی‌های خاصی برخوردار می‌باشد، اما از نظر کارایی عملکرد بهتری نسبت به روش‌های کلاسیک دارد. در این پژوهش، تنها این روش معرفی و با روش‌های کلاسیک مقایسه می‌شود. خوانندگان علاقه‌مند می‌توانند برای مطالعه بیش‌تر به منبع شماره ۷ مراجعه نمایند.

در این مطالعه پرسشنامه خوددرمانی ۵۰۰ نفر از بیماران دیابتی مورد ارزیابی قرار گرفت. از این تعداد، ۴۶ مورد به‌طور ناقص پرسشنامه را تکمیل کرده بودند. در میان این ۴۶ پرسشنامه، تعداد مشاهده‌های گمشده گاهی فقط یک بی‌پاسخی در کل پرسشنامه بود و بیش‌ترین فراوانی را در این میان تعداد ۴ بی‌پاسخی و کم‌تر، به خود اختصاص داده بود. ۴۵۴ شرکت کننده باقی‌مانده به تمام ۲۵ سؤال پرسشنامه خوددرمانی به‌طور کامل پاسخ داده بودند.

برای دست‌یابی به یک دید منطقی در مواجهه با داده‌های گمشده در متغیرهای رتبه‌ای، یک مطالعه روش‌شناختی با استفاده از زیر مجموعه‌ای از شرکت‌کنندگانی که پاسخ‌های کامل داشتند اجرا گردید. برای این منظور با استفاده از روش شبیه‌سازی نسبت به حذف مصنوعی تعدادی از مشاهده‌ها اقدام گردید. برای تولید گمشدگی از ۴ روش مختلف استفاده شد. ابتدا داده‌ها به‌صورت تصادفی با احتمالات ۱۰، ۲۰ و ۳۰ درصد برای تمام سؤالات مفقود شدند. سپس از الگوی گمشدگی با احتمالات نابرابر برای تولید چهارمین مجموعه داده استفاده شد. چگونگی اختصاص این احتمالات مشابه گمشدگی واقعی در ۴۶ پرسشنامه‌ای بود که دارای بی‌پاسخی بودند. سپس ۵ روش جانهی روی هر یک از این ۴ مجموعه داده تولید شده اجرا گردید. علت استفاده از این ۴ روش گمشدگی مختلف به ساختار گمشدگی پرسشنامه‌ها بر می‌گردد. گمشدگی‌ها یا کاملاً تصادفی هستند (MCAR)[1]، یا تصادفی هستند (MAR)[2] و یا غیر تصادفی (MNAR)[3]. علاقه‌مندان برای اطلاع از انواع ساختارهای گمشدگی می‌توانند به منبع شماره ۸ مراجعه نمایند.

روش کار

تعداد ۵۰۰ بیمار دیابتی مراجعه کننده به مرکز دیابت شهرستان کرمانشاه با سابقه حداقل یک‌سال ابتلا به بیماری دیابت وارد مطالعه شدند. پرسشنامه خوددرمانی ویژه بیماران دیابتی که روایی و پایایی آن در مطالعه مسعودی علوی (۱) بررسی شده بود، در اختیار این بیماران قرار گرفت. پرسشنامه شامل ۲۵ سؤال با مقیاس لیکرت ۴ تایی بود، که شرکت‌کنندگان باید نمره‌ای بین ۱ تا ۴ به هر سؤال اختصاص می‌دادند. مجموع نمره‌ها، معیار نمره خوددرمانی در این بیماران قرار گرفت. بنابراین هر شرکت‌کننده نمره‌ای بین ۲۵ تا ۱۰۰ دریافت می‌کرد و نمره بالاتر بیانگر میزان خوددرمانی بیش‌تر در این بیماران بود.

در مجموع ۴۵۴ بیمار به‌طور کامل به پرسشنامه‌ها پاسخ دادند. به منظور تولید مجموعه داده‌های گمشده ساختگی، ابتدا یک عدد تصادفی بین ۰ تا ۱ از توزیع یکنواخت به هر مشاهده اختصاص پیدا کرد (۲). مقدار احتمال اختصاص داده شده به هر مشاهده مبنای حذف آن مشاهده قرار گرفت. ابتدا سه روش گمشدگی کاملاً تصادفی (MCAR) که در آن احتمال گمشدگی به ویژگی‌های افراد مرتبط نیست، شبیه‌سازی شد. به منظور تولید یک مجموعه داده با ۱۰ درصد گمشدگی تمامی مشاهده‌هایی که مقادیر کمتر از ۰/۱۰ دریافت کرده بودند، حذف شدند. سپس این روش برای تولید مجموعه داده‌های با ۲۰ و ۳۰ درصد گمشدگی نیز اجرا گردید.

هم‌چنین یک روش گمشدگی نامتعادل نیز در این مطالعه اجرا گردید. در این روش احتمال گمشدگی هر سؤال متفاوت از سؤال دیگر بود. این احتمال بر اساس میزان گمشدگی واقعی در پرسشنامه‌ها در نظر گرفته شد. به عنوان مثال سؤالاتسؤالات ۲۲ و ۲۳ که به ترتیب دارای ۸ و ۶ گمشدگی بودند، احتمال بیش‌تری برای گمشدگی دریافت کردند و در مقابل سؤالات ۱،۲،۸،۹،۱۰و۱۲ به دلیل عدم گمشدگی احتمالی برابر صفر برای حذف شدن، دریافت نمودند. این روش با عنوان روش نامتعادل در جدول‌ها نمایش داده می‌شود. تمامی تحلیل‌ها و تولید گمشدگی‌های تصنعی در محیط نرم‌افزار SPSS نسخه ۲۰ به انجام رسید.

پنج روش مورد بررسی عبارت بودند از: ۱. میانگین سؤالات؛ ۲. میانگین فردی؛ ۳. نمای فردی؛ ۴. رگرسیون خطی؛ و ۵. الگوریتم EM.

۱) میانگین سؤالات: روش میانگین سؤالات میانگین کلی

[1] Missing Completely at Random
[2] Missing at Random
[3] Missing Not at Random

یک سؤال خاص را محاسبه و جانشین تمام مقادیر گمشده همان سؤال قرار می‌دهد. به عنوان مثال اگر یک شرکت کننده دارای مقدار گمشده در سؤال ۱۰ باشد، میانگین این سؤال برای تمام افرادی که به آن پاسخ داده‌اند، محاسبه و جانشین این مقدار گمشده برای فرد می‌شود.

۲) میانگین فردی: در این روش بر خلاف روش میانگین سؤالات، میانگین نمره خود فرد محاسبه و به جای تمامی مقادیر گمشده فرد قرار می‌گیرد. به عنوان مثال اگر فردی دارای ۲ بی‌پاسخی در سؤالات باشد، میانگین ۲۳ سؤالی که پاسخ داده است محاسبه و به جای این ۲ مقدار گمشده قرار می‌گیرد.

۳) نمای فردی: در این روش پاسخی که بیش‌ترین فراوانی را در میان پاسخ‌های یک فرد به‌دست آورده است، جانشین تمام مقادیر گمشده همان فرد می‌شود. به عنوان مثال اگر گزینه ۱ بیش‌ترین جوابی باشد که یک فرد در پرسشنامه به سؤالات داده است، همین عدد به عنوان جانشینی برای سؤالات بی‌پاسخ انتخاب می‌گردد.

۴) رگرسیون خطی: در این روش ابتدا سؤالی که بیش‌ترین مقدار گمشدگی را دارد، به عنوان متغیر پاسخ و سؤالات دیگر به عنوان متغیر پیشگو در نظر گرفته می‌شوند و پس از برازش خط رگرسیونی مقادیر پیش‌بینی شده جانشین مقادیر گمشده متغیر پاسخ می‌گردند. سپس سؤالی که در رده دوم بیش‌ترین گمشدگی است به عنوان متغیر پاسخ و سؤالات دیگر به عنوان متغیر پیشگو در نظر گرفته می‌شوند. بنابراین نخستین سؤال که دارای بیشترین گمشدگی بود و با مقادیر پیش‌بینی شده جانهی شده بود؛ نیز به عنوان یکی از پیشگوها وارد مدل می‌شود. این روش تا آخرین سؤال ادامه می‌یابد تا در نهایت تمام سؤالات با مقادیر پیش‌بینی شده بر اساس خط رگرسیون برازش داده شده جانشین گردند.

۵) الگوریتم EM: یک روند تکراری مؤثر به منظور محاسبه برآورد حداکثر درست‌نمایی در حضور داده‌های گمشده به حساب می‌آید. هر تکرار الگوریتم شامل دو مرحله می‌باشد: مرحله امید ریاضی (E-step)[1] و مرحله حداکثرسازی (M-step)[2]. در مرحله‌ی امید ریاضی داده‌های گمشده به شرط داده‌های مشاهده شده و برآورد جاری پارامترهای مدل برآورد می‌شوند. در مرحله‌ی حداکثر سازی تابع درست‌نمایی با این فرض که داده‌های گمشده معلوم هستند، حداکثر می‌شود. در حقیقت برآورد داده‌های گمشده از مرحله‌ی امید ریاضی به جای مقادیر گمشده قرار می‌گیرند. با تکرار الگوریتم با توجه به این‌که مقدار درست‌نمایی در هر مرحله افزایش می‌یابد، می‌توان نسبت به هم‌گرایی مطمئن بود (۱۰-۹).

برای هر یک از روش‌های پنج‌گانه بالا، مقدار نمره خوددرمانی ابتدا با مقادیر جانهی شده و سپس با مقادیر واقعی محاسبه شدند. میانگین نمونه و انحراف معیار نمره‌های خوددرمانی با مقدار واقعی مقایسه و ضریب همبستگی اسپیرمن، درصد دسته‌بندی اشتباه و هم‌چنین مقدار آماره‌ی کاپا برای هر روش محاسبه شد. این سه آماره سطح توافق میان روش‌های جانهی شده با مقادیر واقعی را ترسیم کردند. ضریب همبستگی اسپیرمن یک آماره‌ی ناپارامتری بر اساس رتبه مشاهده‌ها برای محاسبه‌ی همبستگی میان متغیرها می‌باشد. آماره کاپا نیز میزان توافق میان نمره خوددرمانی (کم، متوسط و زیاد) را برای مقادیر جانهی شده در برابر نمرات مشاهده شده واقعی محاسبه می‌کند. لاندیس و کوچ مقدار آماره کاپا را به ۵ دسته‌یکم‌تر از ۰/۲ بیان‌گر توافق ضعیف، ۰/۴-۰/۲۱ بیان‌گر توافق کم، ۰/۶-۰/۴۱ توافق متوسط، ۰/۸-۰/۶۱ توافق خوب و بالاتر از ۰/۸۱ توافق عالی تقسیم‌بندی می‌کنند (۱۱).

یافته‌ها

جدول شماره ۱ توزیع حذف شدگی تصادفی را برای هر یک از چهار الگوی گمشدگی (P=۰/۱، P=۰/۲، P=۰/۳ و نامتعادل) نشان می‌دهد. همان‌طور که مشاهده می‌شود با افزایش درصد گمشدگی متوسط تعداد مقادیر گمشده افزایش می‌یابد. در احتمال ۱۰ درصد، بیش‌تر شرکت‌کنندگان دارای یک مقدار گمشدگی ساختگی هستند. هنگامی که گمشدگی به ۳۰ درصد افزایش می‌یابد، بیش‌تر شرکت‌کنندگان بین ۹-۷ مشاهده حذف شده تصادفی دارند.

جدول‌های شماره ۵-۲ میانگین، انحراف معیار، ضریب

[1] Expectation step

[2] Maximisation step

همبستگی اسپیرمن، درصد دسته‌بندی اشتباه و آماره‌ی کاپا را برای هر یک از روش‌های جانهی نشان می‌دهد. وقتی داده‌ها دارای درصد گمشدگی کم‌تری (P=۰/۱) هستند (جدول شماره ۲)، مقدار آماره کاپا برای الگوریتم EM بیش‌تر از ۰/۸۱ است که نشانه‌ی توافق تقریباً کامل با مشاهده‌های واقعی است. در همین سطح، روش میانگین سؤالات کم‌ترین مقدار آماره کاپا را نشان می‌دهد. میانگین روش‌های نمای فردی (۴۹/۱۹) و میانگین سؤالات (۴۹/۹۱) به‌طور معنی‌داری با میانگین واقعی (۴۹/۶۵) تفاوت دارد؛ در حالی‌که روش رگرسیون خطی نزدیک‌ترین انحراف معیار (۵/۴۲۳) را با مقدار واقعی آن (۵/۴۶۶) نشان می‌دهد. روش الگوریتم EM بالاترین مقدار آماره‌ی کاپا (۰/۸۸۶) و کم‌ترین درصد دسته‌بندی اشتباه (۵/۵ درصد) را نشان می‌دهد.

شکل شماره ۱ همبستگی میان مقادیر پیش‌بینی شده و مشاهده شده را وقتی احتمال گمشدگی در داده‌ها ۲۰ درصد است؛ را نمایش می‌دهد. روش‌های الگوریتم EM و میانگین فردی خوشه‌های باریک‌تری را نشان می‌دهد. پراکندگی در رگرسیون خطی و روش میانگین سؤالات بیش‌تر از سایر روش‌ها می‌باشد.

با افزایش گمشدگی اطلاعات (P=۰/۲ و P=۰/۳) روش الگوریتم EM هم‌چنان از چهار روش دیگر دقت بیش‌تری نشان می‌دهد. مقدار آماره‌ی کاپا با افزایش احتمال گمشدگی هم‌چنان بالا بوده که نشان دهنده مناسب بودن این روش حتی در خصوص گمشدگی‌های اساسی می‌باشد. مقدار آماره کاپا برای این روش با مقدار ۰/۶۹۷ هنوز در بازه‌ی توافق خوب قرار دارد. با این حال درصد دسته‌بندی اشتباه برای وضع خوددرمانی نسبت به حالت اول تقریباً حدود ۱۰ درصد افزایش یافته است. روش میانگین فردی نیز عملکرد منطقی خوبی را با افزایش احتمال گمشدگی نشان می‌دهد.

روش الگوریتم EM و میانگین فردی درصد مشابهی از دسته‌بندی اشتباه و ثبات در آماره‌ی کاپا را با افزایش گم‌شدگی از ۰/۱ به ۰/۳ نشان می‌دهند. انحراف معیار هر دو روش تقریباً متفاوت از انحراف معیار داده‌های واقعی است.

مقدار آماره کاپا برای روش رگرسیونی در رتبه قابل قبولی قرار ندارد. به‌ویژه برای الگوی ۲۰ درصد گمشدگی این مقدار تنها ۰/۶۳۵ محاسبه شده است.

روش نامتعادل نتایجی مشابه احتمال گمشدگی ۱۰ درصد ایجاد می‌کند. الگوریتم EM هم‌چنان در این روش نیز از قدرت بیش‌تری در برآورد مقادیر گمشده برخوردار می‌باشد.

جدول شماره ۱- فراوانی تعداد گمشدگی های تصنعی برای هر مشاهده تحت گمشدگی های کاملاً تصادفی و روش نامتعادل

سبک گمشدگی داده‌ها	p=۰/۱		P=۰/۲		P=۰/۳		روش نامتعادل	
کل پاسخ‌های گمشده	N	درصد کل	N	درصد کل	N	درصد کل	N	درصد کل
۰	۵۳	۱۱/۷	۵	۱/۱	۱	۰/۲	۱۵۹	۳۵
۱	۱۰۸	۲۳/۸	۱۱	۲/۴	۰	۰	۱۶۰	۳۵/۲
۲	۸۶	۱۸/۹	۲۹	۶/۴	۴	۰/۹	۹۸	۲۱/۶
۳	۸۲	۱۸/۱	۵۶	۱۲/۳	۱۵	۳/۳	۳۳	۷/۳
۴	۶۰	۱۳/۲	۹۵	۲۰/۹	۳۰	۶/۶	۳	۰/۷
۵	۳۴	۷/۵	۹۴	۲۰/۷	۵۱	۱۱/۲	۱	۰/۲
۶	۲۴	۵/۳	۶۳	۱۳/۹	۶۰	۱۳/۲	۰	۰
۷	۵	۱/۱	۵۴	۱۱/۹	۷۵	۱۶/۵	۰	۰
۸	۱	۰/۲	۲۷	۵/۹	۶۹	۱۵/۲	۰	۰
۹	۱	۰/۲	۱۱	۲/۴	۶۸	۱۵	۰	۰
۱۰	۰	۰	۶	۱/۳	۴۱	۹	۰	۰
۱۱	۰	۰	۳	۰/۷	۱۸	۴	۰	۰
۱۲	۰	۰	۰	۰	۱۱	۲/۴	۰	۰
۱۳	۰	۰	۰	۰	۶	۱/۳	۰	۰
۱۴	۰	۰	۰	۰	۴	۰/۹	۰	۰
۱۵	۰	۰	۰	۰	۰	۰	۰	۰
۱۶	۰	۰	۰	۰	۱	۰/۲	۰	۰
۱۷	۰	۰	۰	۰	۰	۰	۰	۰

جدول شماره ۲- آماره‌های تشخیصی در الگوی گمشدگی ۱۰ درصد

سبک گمشدگی داده‌ها	روش	میانگین	انحراف معیار	ضریب اسپیرمن	دسته‌بندی اشتباه	کاپا
P=۰/۱ N=۱۱۳۲ μ =۴۴/۹۱ σ =۶/۰۶	میانگین سؤالات	۴۹/۹۱	۵/۶۸۱	۰/۹۳۰	(۸/۸)۴۰	۰/۸۱۷
	میانگین فردی	۴۹/۷۱	۵/۰۳۵	۰/۹۵۶	(۷/۵)۳۴	۰/۸۴۱
	نمای فردی	۴۹/۱۹	۵/۰۶۷	۰/۹۴۶	(۷/۹)۳۶	۰/۸۲۵
	رگرسیون خطی	۴۹/۷۱	۵/۴۲۳	۰/۹۴۳	(۷/۹)۳۶	۰/۸۳۷
	الگوریتم EM	۴۹/۶۷	۵/۲۷۱	۰/۹۶۴	(۵/۵)۲۵	۰/۸۸۶

جدول شماره ۳- آماره‌های تشخیصی در الگوی گمشدگی ۲۰ درصد

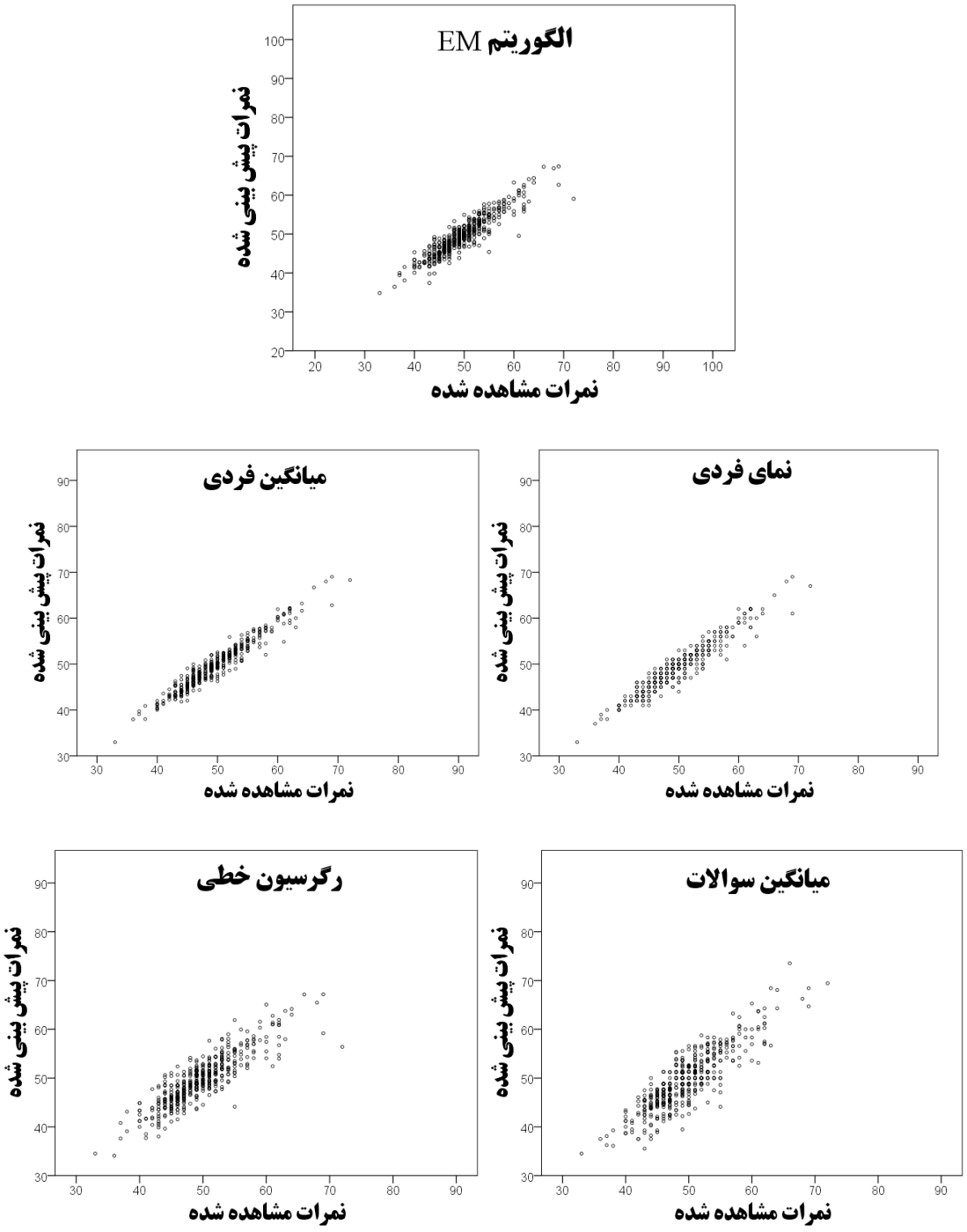
سبک گمشدگی داده‌ها	روش	میانگین	انحراف معیار	ضریب اسپیرمن	دسته‌بندی اشتباه	کاپا
p=۰/۲ N=۲۲۵۱ μ =۳۹/۷۳ σ =۶/۱۵۶	میانگین سؤالات	۴۹/۵۹	۵/۹۹۴	۰/۸۳۶	(۱۸/۵)۸۴	۰/۶۰۸
	میانگین فردی	۴۹/۶۰	۴/۴۸۲	۰/۹۰۷	(۱۳/۲)۶۰	۰/۷۲۱
	نمای فردی	۴۸/۵۲	۴/۵۵۸	۰/۸۸۷	(۱۲/۱)۵۵	۰/۷۲۹
	رگرسیون خطی	۴۹/۶۶	۵/۱۵۹	۰/۸۶۶	(۱۷/۸)۸۱	۰/۶۳۵
	الگوریتم EM	۴۹/۵۹	۴/۹۸۹	۰/۹۰۸	(۱۳/۲)۶۰	۰/۷۲۶

جدول شماره ٤- آماره‌های تشخیصی در الگوی گمشدگی ۳۰ درصد

سبک گمشدگی داده‌ها	روش	میانگین	انحراف معیار	ضریب اسپیرمن	دسته‌بندی اشتباه	کاپا
P=۰/۳ N=۳۳۶۷ μ =۳۴/۸۳ σ =۶/۴۳۲	میانگین سؤالات	۴۹/۵۵	۶/۵۰۱	۰/۷۷۰	(۲۰/۷)۹۴	۰/۵۶۷
	میانگین فردی	۴۹/۵۹	۴/۱۳۴	۰/۸۷۷	(۱۴/۷)۶۷	۰/۶۸۷
	نمای فردی	۴۷/۹۱	۴/۲۶۳	۰/۸۵۴	(۱۶/۷)۷۶	۰/۶۱۶
	رگرسیون خطی	۴۹/۴۸	۵/۰۱۲	۰/۸۱۶	(۱۷/۴)۷۹	۰/۶۴۰
	الگوریتم EM	۴۹/۵۸	۴/۸۳۱	۰/۸۷۴	(۱۴/۵)۶۶	۰/۶۹۷

جدول شماره ۵- آماره‌های تشخیصی در الگوی گمشدگی نامتعادل

سبک گمشدگی داده‌ها	روش	میانگین	انحراف معیار	ضریب اسپیرمن	دسته‌بندی اشتباه	کاپا
روش نامتعادل N=۴۷۲ μ =۴۷/۶۶ σ =۵۵۹۵	میانگین سؤالات	۴۹/۷۴	۵/۵۷۹	۰/۹۶۳	(۷/۷)۳۵	۰/۸۳۷
	میانگین فردی	۴۹/۶۶	۵/۲۹۹	۰/۹۸۰	(۶/۴)۲۹	۰/۸۶۷
	نمای فردی	۴۹/۳۴	۵/۳۱۱	۰/۹۷۸	(۵/۷)۲۶	۰/۸۷۷
	رگرسیون خطی	۴۹/۵۹	۵/۴۳۲	۰/۹۶۸	(۷/۳)۳۳	۰/۸۴۹
	الگوریتم EM	۴۹/۶۲	۵/۳۸۸	۰/۹۸۵	(۴/۶)۲۱	۰/۹۰۳



شکل شماره ۱- نمودار پراکندگی نمره‌های مشاهده شده در مقابل نمرات برآورد شده با روش‌های الگوریتم EM، میانگین فردی، رگرسیون خطی، میانگین و میانه سؤالات

بحث

بی‌پاسخی در مطالعه‌های مبتنی بر پرسشنامه امری اجتناب‌ناپذیر است. افزایش بی‌پاسخی‌ها، بر تحلیل‌ها تأثیر و توان مطالعه را کاهش می‌دهد. بنابراین پژوهشگر ناگزیر است به‌دنبال روش‌هایی برای حل مسأله گمشدگی باشد. در این مطالعه موردی به زبانی ساده به معرفی و مقایسه جانهی داده‌های گمشده با بهره از روش‌های کلاسیک و روش مدرن الگوریتم EM هنگام مواجهه با

داده‌های پرسشنامه‌ای پرداخته شد. هم‌چنین با ایجاد گمشدگی‌های تصادفی و غیر تصادفی که در شرایط مختلف می‌تواند رخ دهد، این مقایسه به چالش کشیده شد. روش الگوریتم EM دقیق‌ترین روش در میان ۴ روش دیگر برای تمام الگوهای گمشدگی مختلف بود. با این حال روش میانگین فردی که روش جانهی ساده‌تری است، عملکرد قابل قبولی داشت. در روش نامتعادل نیز با وجود الگوی گمشدگی متفاوت الگوریتم EM رفتار بهتری را از خود نشان داد.

اگرچه الگوریتم EM احتمالاً دقیق‌ترین و معتبرترین روش در این مطالعه بود، اما معایبی نیز دارد. این روش پیچیده و نیازمند روش‌های پیشرفته آماری است که احتمالاً برای بسیاری از خوانندگان و پژوهشگران ناآشنا است.

روش رگرسیون خطی رفتاری مشابه تکنیک جانهی چندگانه دارد که به جای چند تکرار فقط یک تکرار در آن اتفاق می‌افتد. این روش بر خلاف انتظار رفتار قابل قبولی در برآورد مقادیر گمشده از خود نشان نداد. یکی از دلایل آن می‌تواند نوع متغیرهای اندازه‌گیری شده در این مطالعه باشد.

روش جانهی میانگین فردی روشی ساده‌تر و در نتیجه قابل فهم‌تر برای بسیاری از خوانندگان پزشکی خواهد بود. در واقع این روش یک روش شهودی برای جانهی مقادیر به حساب می‌آید. فرض اساسی این است که افراد پاسخ‌های مشابهی در سراسر پرسشنامه دارند که یک فرض منطقی برای پرسشنامه‌های دارای پاسخ رتبه‌ای همانند پرسشنامه خوددرمانی محسوب می‌شود (۱۰). در این مطالعه، میانگین فردی ضرایب همبستگی قابل اعتباری فراهم نمود. با این حال، این روش معمولاً هنگامی که گمشدگی‌ها از نوع تصادفی باشد و درصد گمشدگی کمی در داده‌ها اتفاق بیافتد؛ از کارایی لازم برخوردار خواهد بود.

نمودار پراکندگی برای میانگین سؤالات که در شکل شماره ۱ ترسیم شده است، حالتی از بیش‌برآوردی نمره‌ها در بیماران با نمره‌های خوددرمانی کم‌تر و کم‌برآوردی نمره‌ها در بیماران با نمره‌های بالاتر را نشان می‌دهد. این ترکیب منجر به چرخش مقادیر مشاهده شده در نمودار پراکندگی شده است. این چرخش در دیگر نمودارها مشاهده نمی‌شود.

اگرچه نتایج به‌دست آمده در این مطالعه به منظور دست‌یابی به یک راه‌حل خوب نسبتاً شفاف می‌باشد، با این حال این یافته‌ها ممکن است در مورد الگوهای گمشدگی دیگر قابل کاربرد نباشد و یک مجموعه داده جدید و یا پرسشنامه‌ای دیگر ممکن است به نتایج متفاوتی منجر گردد.

یافته‌های این مطالعه با نتایج جی مل (۱۲) و هاوتورن و همکاران (۱۳) که عملکرد روش‌های جانهی مختلف را روی پرسشنامه‌ها ارزیابی کردند، مشابه بود. هر دوی این مطالعه‌ها نشان دادند که اگرچه روش‌های جانهی پیچیده مانند روش جانهی هات- دک دارای مزایایی است، با این حال روش‌های جانهی تک مقداری هم‌چون میانگین فردی عملکرد مناسبی از خود نشان می‌دهد، اما هنگامی که گمشدگی از نوع غیر تصادفی است و درصد بالایی از مشاهده‌ها گمشده هستند، نتایج حاصل از این روش‌ها قابل اعتماد نخواهد بود. نویسندگان دیگر اظهار داشتند که اگرچه روش‌هایی چون جانهی میانگین از نظر ریاضی ساده‌تر هستند، با این حال منجر به کم‌برآوردی واریانس در میان داده‌ها می‌شود و بهتر است از روش‌هایی چون جانهی چندگانه استفاده شود (۲،۹،۱۱). فارکلو و سلا عملکرد قوی روش جانهی ساده که مشابه روش میانگین فردی در این مطالعه بود را نشان دادند (۱۴). فایر و همکاران (۱۵) خاطر نشان می‌کنند که اگرچه روش‌های جانهی ساده اغلب عملکرد خوبی دارند، با این حال در خصوص استفاده از آن در حالت گسترده باید احتیاط کرد. آن‌ها چک‌لیستی مفید برای پژوهشگران فراهم آوردند که باید هنگام استفاده از روش‌های جانهی ساده در نظر بگیرند.

این مطالعه نشان داد استفاده از روش‌های کلاسیک هم‌چون میانگین فردی هنگام مواجهه با گمشدگی در مطالعه‌هایی که از ابزار پرسشنامه استفاده می‌کنند، می‌تواند یک روش قابل اعتماد برای جانهی مقادیر گمشده به شمار آید. اگرچه در این مطالعه روش میانگین فردی با گمشدگی‌های مختلف و حتی زمانی که گمشدگی در پرسشنامه‌ها به‌صورت غیر تصادفی اتفاق افتاده بود؛ نتایجی مشابه الگوریتم EM ارایه نمود، با این حال استفاده از این روش تنها زمانی که گمشدگی از نوع کاملاً تصادفی باشد و درصد گمشدگی‌ها کم باشد؛ مشابه روش‌های پیشرفته خواهد بود. اما همان گونه که گمشدگی‌ها در عمل معمولاً از نوع غیر تصادفی هستند، بنابراین استفاده از روش‌های کلاسیک در این شرایط باید با احتیاط و تأمل بیش‌تر صورت گیرد (۱۶).

نتیجه‌گیری

روش الگوریتم EM باید در شرایطی که داده‌ها به‌صورت کاملاً تصادفی گمشده‌اند؛ مورد استفاده قرار گیرد (۷). در این مطالعه

روش الگوریتم EM نه تنها در حالت گمشدگی تصادفی که در حالت گمشدگی با احتمالات نابرابر نیز عملکرد خوبی از خود نشان داد.

اگرچه در این مطالعه الگوریتم EM از توانمندی بیش‌تری در برآورد مقادیر گمشده برخوردار بود؛ با این وجود چون روش میانگین فردی تعادل خوبی میان سادگی و دقت فراهم می‌آورد، بنابراین می‌تواند به عنوان بهینه‌ترین روش در این مطالعه معرفی شود. روش‌های کلاسیک جانهی داده‌ها، در مطالعه‌های پرسشنامه‌ای با گمشدگی‌های کاملاً تصادفی می‌تواند گزینه‌ای جذاب برای پژوهشگران بالینی به شمار آید.

منابع

1. Masoudi AN, Alami L, Taefi S, Sadafi Z. Self treatment in diabetes mellitus in Kashan. Iranian Journal of Endocrinology and Metabolism (JEM). 2010; 12: 237-242.
2. Shrive FM, Stuart H, Quan H, Ghali WA. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC Medical Research Methodology. 2006; 6: 1-10.
3. Hashemian A-H, Afshari-Safavi A, Rezaei M, Payandeh M, Golpayegani M-R, Fallah-Pakdel S. Detecting the determinants of cardiac and hepatic iron overload in patients with thalassemia major using a generalized estimating equations method. Journal of Mazandaran University of Medical Sciences (JMUMS). 2014; 23:175-181.
4. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005; 61: 962-73.
5. Hogan JW. Comments on: Missing data methods in longitudinal studies: a review. Test. 2009; 18: 59-64.
6. Oliveira P, Gomes L. Interpolation of signals with missing data using Principal Component Analysis. Multidimensional Systems and Signal Processing. 2010; 21: 25-43.
7. Mclachlan GJ, Krishnan T. The EM Algorithm and Extensions. 2nd ed. John Wiley and Sons Inc. 2008: 77-84.
8. Zayeri F, Akbarzadeh Baghban AR, Kazemzadeh M, Yaseri M, Abbasi AM. Different Types of Missing in Longitudinal Data and the Likelihood-base Methods Applied in their Analysis. Journal of Ilam University of Medical Sciences. 2013, 20: 208-22.
9. Ghomrawi HM, Mandl LA, Rutledge J, Alexiades MM, Mazumdar M. Is there a role for expectation maximization imputation in addressing missing data in research using WOMAC questionnaire? Comparison to the standard mean approach and a tutorial. BMC musculoskeletal disorders. 2011; 12: 1-7.
10. Downey RG, King CV. Missing data in Likert ratings: A comparison of replacement methods. The Journal of General Psychology. 1998; 125: 175-91.
11. Fay RE. When are inferences from multiple imputation valid.Proceedings of the Survey Research Methods Section of the American Statistical Association. 1992; 81: 227-32.
12. Gmel G. Imputation of missing values in the case of a multiple item instrument measuring alcohol consumption. Statistics in medicine. 2001; 20: 2369-81.
13. Hawthorne G, Elliott P. Imputing crosssectional missing data: comparison of common techniques. Australian and New Zealand Journal of Psychiatry. 2005; 39: 583-90.
14. Fairclough D, Cella D. Functional Assessment of Cancer Therapy (FACT-G): Non-response to individual questions. Quality of Life Research. 1996; 5: 321-9.
15. Fayers PM, Curran D, Machin D. Incomplete quality of life data in randomized trials: missing items. Statistics in medicine. 1998;17: 679-96.
16. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons; 1987: 15-22.

Iranian Journal of Epidemiology 2015; 11(3): 43-51.

Original Article

Comparison of EM Algorithm and Standard Imputation Methods for Missing Data: A Questionnaire Study on Diabetic Patients

Afshari Safavi A[1], Kazemzadeh Gharechobogh H[2], Rezaei M[3]

1- PhD Student of Biostatistics, Student Research Committee, Isfahan University of Medical Sciences, Isfahan, Iran
2- MSc of Statistics, Social Security Organization, Tehran, Iran
3- Department of Biostatistics and Epidemiology, Social Development and Health Promotion Research Center, Kermanshah University of Medical Sciences, Kermanshah, Iran
Corresponding author: Kazemzadeh H, kazemzadeh_hk@yahoo.com

Background and Objectives: Missing data is a big challenge in the research. According to the type of the study and of the variables, different ways have been proposed to work with these data. This study compared five popular imputation approaches in addressing missing data in the questionnaires.

Methods: In this study, 500 questionnaires were used for self-medication in diabetic patients. Missing in the observations was artificially generated by random selection of questions and then deleting them. Five imputation ways included: 1) the mean of the questions, 2) the mean of the person, 3) the mode of the person, 4) linear regression, and 5) EM algorithm. For each method, the mean and standard deviation were compared with imputation. The Spearman correlation coefficient, the percentage of incorrectly classified and kappa statistic were also calculated.

Results: A kappa higher than 0.81 represented almost perfect agreement at 10% missingness. The EM algorithm showed the highest level of agreement with the results of actual data with a Kappa of 0.886. With increasing missingness to 30%, the EM algorithm and the mean of the person showed a rather similar agreement with a Kappa of 0.697 and 0.687, respectively.

Conclusion: In this study, the EM algorithm was the most accurate method for handling missing data in all models. The mean of the person method is easy for handling missing data, especially for most non statisticians.

Keywords: Algorithm EM, Missing data, Diabetes, Self-treatment, Kappa statistics, Regression