مجله تخصصی اپیدمیولوژی ایران
https://irje.tums.ac.ir/
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران
IrEA انجمن علمی اپیدمیولوژیست های ایران

مقاله آموزشی

داده‌های گم‌شده (Missing Data) در پژوهش‌های علوم پزشکی: مروری کاربردی و مصور برای پژوهشگران و دانشجویان

پریسا امجدی زین حاجلو[۱]، محمد حیدری[۲]

۱- کارشناس ارشد اپیدمیولوژی، گروه اپیدمیولوژی و آمار زیستی، دانشکده پزشکی، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران
۲- استادیار اپیدمیولوژی، گروه اپیدمیولوژی و آمار زیستی، دانشکده پزشکی، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران

DOI:

اطلاعات مقاله

تاریخ دریافت
۱۴۰۴/۰۶/۱۵
تاریخ پذیرش
۱۴۰۴/۱۰/۱۵
نویسنده رابط
محمد حیدری
ایمیل نویسنده رابط
heidari.m@umsu.ac.ir
نشانی نویسنده رابط
ارومیه، کیلومتر 11 جاده سرو، پردیس نازلو، دانشکده پزشکی، گروه اپیدمیولوژی و آمار

واژگان کلیدی: داده های گم‌شده، الگوهای گمشدگی، MCAR، MAR، MNAR، جای‌گذاری چندگانه

چکیده

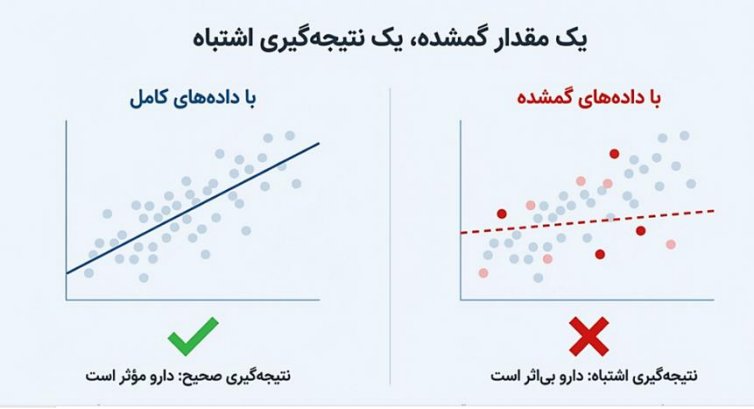
داده‌های گم‌شده یکی از چالش‌های مهم در پژوهش‌های علوم پزشکی و اپیدمیولوژی هستند و در صورت مدیریت نامناسب می‌توانند منجر به سوگیری، کاهش توان آماری و برداشت‌های نادرست شوند. با وجود اهمیت این موضوع، منابع فارسی جامع و درعین‌حال کاربردی در این زمینه محدود است. این مقاله با رویکرد آموزشی، مروری روشن و منسجم بر مفاهیم پایه‌ای داده‌های گم‌شده از جمله تعریف‌ها، الگوهای بروز (تک‌متغیره و چندمتغیره) و سه سازوکار اصلی گم‌شدگی شامل MCAR، MAR و MNAR ارائه می‌کند. همچنین طیفی از روش‌های رایج در مدیریت داده‌های گم‌شده، از حذف موارد گم‌شده تا جای‌گذاری‌های ساده و رویکردهای پیشرفته‌تر مانند جای‌گذاری چندگانه و روش‌های مبتنی بر درست-نمایی نظیر الگوریتم EM و MLE به‌طور خلاصه و قابل‌فهم ارائه شده و برای تقویت درک مفهومی و کاربردی‌سازی مطالب، از مثال‌های روشن و مصورسازی‌های آموزشی بهره گرفته شده است. هدف نهایی مقاله فراهم کردن چارچوبی عملی برای پژوهشگران و دانشجویان است تا بتوانند در طراحی و تحلیل پژوهش‌های خود، رویکرد مناسب را برای مواجهه با داده‌های گم‌شده برگزینند و از بروز خطاهای تحلیلی پیشگیری کنند.

مقدمه

وجود داده‌های گم‌شده به عنوان امری فراگیر و اجتناب ناپذیر تقریباً در تمامی مطالعات علمی، از جمله نظر سنجی‌ها و آزمایش‌هایی با طراحی دقیق مشاهده می‌شود. این پدیده به‌خصوص در پژوهش‌های طولی، که مستلزم انجام اندازه‌گیری‌های مکرر در نقاط زمانی مختلف بر روی یک گروه نمونه ثابت است، چالش برانگیزتر می‌شود و می‌تواند اثرات منفی عمیقی بر صحت و قابلیت استنتاج نتایج داشته باشد (۱, ۲). وجود داده‌های گم‌شده می‌تواند به کاهش قابل توجه قدرت آماری تحلیل‌ها و افزایش احتمال خطای نوع اول (رد فرضیه صفر صحیح) منجر شود، افزون بر این، داده‌های گم‌شده موجب ایجاد سوگیری در برآورد پارامترها شده و در نهایت می‌توانند به نتایج تحریف‌شده، نادقیق یا گمراه‌کننده بینجامند (۳). اگرچه مبانی نظری رویکرد به داده‌های گم‌شده پیش‌تر توسط پژوهش‌های بنیادین دمپستر و همکاران (۱۹۷۷)، هکمن (۱۹۷۹) و روبین (۱۹۷۶) نهایی شده بود اما توجه جدی و گسترده به این مسئله از دهه ۱۹۸۷ آغاز شد. نقطه عطف در این حوزه، انتشار دو کتاب مهم بود که نقش محوری در توسعه روش شناختی این علم ایفا

کردند. یکی از این منابع، کتاب «تحلیل آماری با داده‌های گم‌شده» اثر لیل و روبین است که در سال ۱۹۸۷ منتشر شد و نسخه دوم آن در سال ۲۰۰۲ به نشر رسید (۱).

مطالعات در ایران نشان داده‌اند که در سامانه‌های پرونده الکترونیک سلامت، فقدان ثبت کامل و استاندارد داده‌ها همچنان یک مشکل عمده است (٤). بنابراین، گمشدگی اطلاعات در پژوهش‌های داخلی واقعاً مسئله‌ای جدی است که عدم رسیدگی صحیح به این داده‌های گم‌شده می‌تواند به سوگیری‌های آماری خطرناکی منجر شود که نتیجه‌گیری بالینی را کاملاً وارونه می‌سازد. همان‌طور که در شکل ۱ تجسم شده است:



شکل شماره ۱- مقایسه نتایج تحلیل رگرسیونی در داده‌های کامل و داده‌های دارای گمشدگی

- با داده‌های کامل (سمت چپ)، تحلیل رگرسیون ممکن است به‌درستی نشان دهد که یک دارو "مؤثر" است (شیب مثبت قوی).
- با همان داده‌ها، اما پس از گم شدن سوگیرانه مقادیر کلیدی (سمت راست)، خط رگرسیون می‌تواند کاملاً مسطح یا معکوس شود و ما را به نتیجه‌گیری اشتباه "دارو بی‌اثر است" برساند.

این انحراف آماری، استنتاج‌های بالینی زیانباری را به دنبال خواهد داشت. لذا، شناسایی و مدیریت دقیق داده‌های گم‌شده پیش از آغاز تحلیل آماری رسمی، از اهمیت بالایی برخوردار است. این مقاله علاوه بر مرور مفاهیم بنیادی داده‌های گم‌شده، یک مثال عددی بومی را برای مقایسه عملی چهار رویکرد متداول حذف کامل، جای‌گذاری با میانگین، جای‌گذاری تک مقداری و جای-گذاری چندگانه ارائه می‌کند. این تحلیل‌های انجام شده با خروجی‌های نرم افزار و تفسیر نتایج در منابع موجود فارسی ارائه نشده‌اند و لذا ارزش افزوده کاربردی مهمی برای پژوهشگران کشور فراهم می‌کنند.

روش

این مقاله یک نوشته آموزشی و مرور مفهومی است که با هدف تبیین اصول، الگوها و روش‌های مدیریت داده‌های گم‌شده تدوین شده است. برای گردآوری محتوای علمی، جستجوی هدفمند در منابع معتبر شامل PubMed، Scopus، Web of Science و Google Scholar انجام شد. کلمات کلیدی Missing data، MCAR[1]، MAR[2]، MNAR[3]، multiple imputation، EM algorithm و data management در بازه زمانی اسفند ۱٤۰۳ تا خرداد ۱٤۰٤ مورد استفاده قرار گرفت. منابع بنیادی کلاسیک از نویسندگان اصلی حوزه Donald B. Rubin، Roderick J. A. Little، James W. Graham و Paul D. Allison نیز به‌صورت هدفمند انتخاب و مرور شدند (۱, ۵-۷). معیار انتخاب منابع، اعتبار علمی، ارتباط مستقیم با موضوع آموزشی، و قابلیت انتقال مفاهیم به زبان ساده بود. این مقاله با هدف آموزشی تدوین شده و شامل مرور جامع نظام‌مند نیست.

[1] Missing completely at random
[2] Missing at Random
[3] Missing Not At Random

تعریف داده‌های گم‌شده

داده‌های گم‌شده به مقادیر داده‌ای اطلاق می‌شوند که برای یک متغیر مشخص در یک مشاهده خاص، ثبت نشده و در مجموعه داده، جای خالی باقی مانده است (۲). در ادبیات آماری، اصطلاحات متعددی از جمله: مقادیر گم-شده[1]، داده‌های گم‌شده[2]، داده‌های ناقص[3] و بی پاسخ[4] برای اشاره به این مفهوم به کار می‌رود (۸). وقوع داده‌های گم‌شده می‌تواند در شرایط مختلفی رخ دهد؛ برای نمونه، ممکن است در فرآیند نظرسنجی، پاسخی به یک سؤال خاص ثبت نشود یا در آزمایشگاه‌های پزشکی، نمونه خون هنگام حمل به دلیل خطای انسانی یا به صورت تصادفی از دست برود. علاوه بر این، امتناع شرکت‌کنندگان از ادامه همکاری در طول یک پژوهش نیز منجر به ایجاد داده‌های گم‌شده می‌گردد. در نتیجه عدم مدیریت داده‌های گم‌شده، منجر به کاهش اعتبار و قابلیت اعتماد نتایج آزمایش‌ها و پژوهش‌ها شده، و در نهایت، منجر به استنتاج‌های نادرست می‌شود (۹).

چرا داده‌ها گم می‌شوند؟

بروز داده‌های گم‌شده در پژوهش‌های علوم پزشکی دلایل متعددی دارد. بخشی از این مشکل ناشی از خطاهای انسانی در مرحله ثبت و وارد کردن داده‌ها می‌باشد که به-ویژه در مطالعاتی با حجم داده بالا، کیفیت اطلاعات را کاهش می‌دهد. علاوه بر این، نقص تجهیزات یا اختلال در نرم‌افزار از عوامل فنی رایج در گمشدگی داده‌ها هستند. در مطالعات مبتنی بر پرسش‌نامه، عدم پاسخگویی شرکت‌کنندگان به سؤالات خاص (به‌ویژه موارد حساس یا مرتبط با حریم خصوصی) می‌تواند منجر به گمشدگی شود. در مطالعات طولی نیز انصراف یا ریزش شرکت‌کنندگان به دلایل شخصی، تغییر محل زندگی یا مشکلات پیگیری، بخش قابل توجهی از گمشدگی داده‌ها را رقم می‌زند. همچنین ماهیت و طولانی بودن دوره پیگیری، احتمال از دست رفتن داده‌ها را افزایش می‌دهد؛ امری که در مطالعات کوهورت و پیگیری همواره چالش برانگیز بوده است. در نهایت محدودیت‌های زمانی، مالی و نیروی انسانی نیز می‌توانند مانع جمع آوری کامل داده-ها شوند و به گمشدگی یا کاهش کیفیت اطلاعات منجر گردند (۱۰).

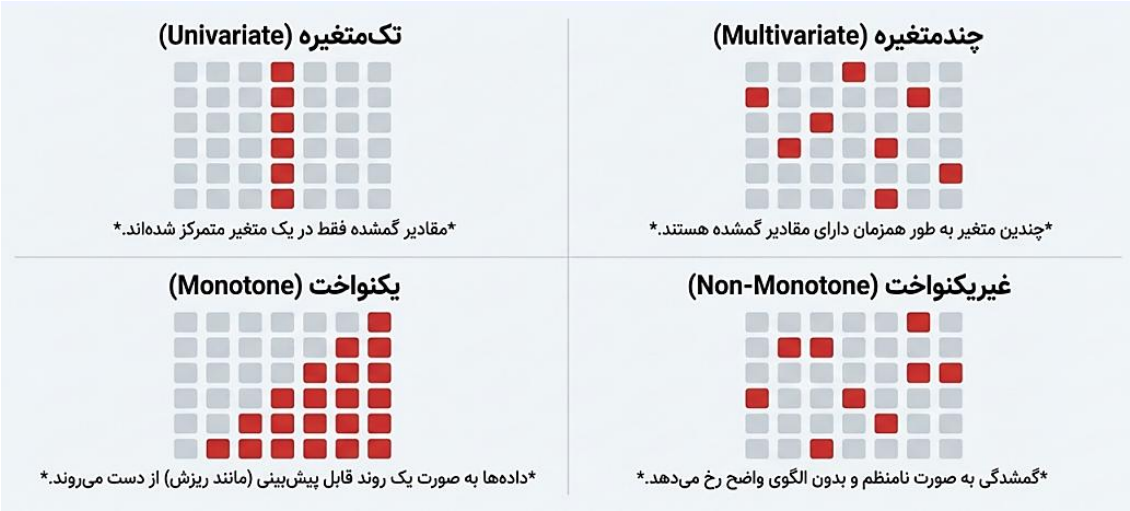
الگوهای داده‌های گم‌شده

الگوی داده‌های گم‌شده بیانگر نحوه توزیع و پراکندگی مقادیر گم‌شده در یک مجموعه داده است که تفکیک دقیق میان آن‌ها برای انتخاب روش تحلیلی مناسب ضروری می‌باشد (شکل ۲). این الگوها شامل: ۱-الگوی گم شدگی تک متغیره[5]: زمانی رخ می‌دهد که مقادیر از دست رفته فقط در یک متغیر متمرکز باشند؛ مانند حالتی که پاسخ‌دهندگان تنها به یک سؤال پرسش‌نامه پاسخ نمی‌دهند. این الگو نسبتاً کمتر مشاهده می‌شود. ۲- الگوی گم شدگی چند متغیره[6]: در این حالت، چندین متغیر هم‌زمان دارای مقادیر گم‌شده‌اند، این الگو در پژوهش‌های واقعی بسیار شایع است و معمولاً نیازمند روش‌های مدیریت چندمتغیره می‌باشد. ۳- الگوی گم شدگی یکنواخت[7]: داده‌ها به صورت سیستماتیک و در یک روند قابل پیش‌بینی از دست می‌روند. به عنوان مثال، در یک مطالعه پیگیری پنج‌ساله، اگر در سال چهارم، فرد در دسترس نباشد، داده‌های مربوط به آن سال از دست می‌رود و این روند ممکن است در سال‌های بعد نیز ادامه یابد. ٤- الگوی گم شدگی غیر یکنواخت[8]: داده‌ها به صورت نامنظم و بدون الگوی قابل پیش‌بینی گم می‌شوند. این نوع الگو بیشتر در مطالعاتی اتفاق می‌افتد که شامل اندازه‌گیری‌های مکرر[9] هستند. برای نمونه، در یک پیگیری ماهانه برای یک بیماری مزمن، ممکن است فرد در ماه‌های اول و دوم در دسترس باشد، ولی ماه سوم، به صورت غیر قابل پیش‌بینی گم شود و در ماه‌های

[1] Missing values
[2] Missing Data
[3] Incomplete Data
[4] non-response

[5] Univariate
[6] Multi-variate
[7] Monotone
[8] Non Monotone
[9] Repeated measures

بعد دوباره در دسترس باشد. تحلیل این الگو، به روش‌های انعطاف‌پذیرتری نیاز دارد.



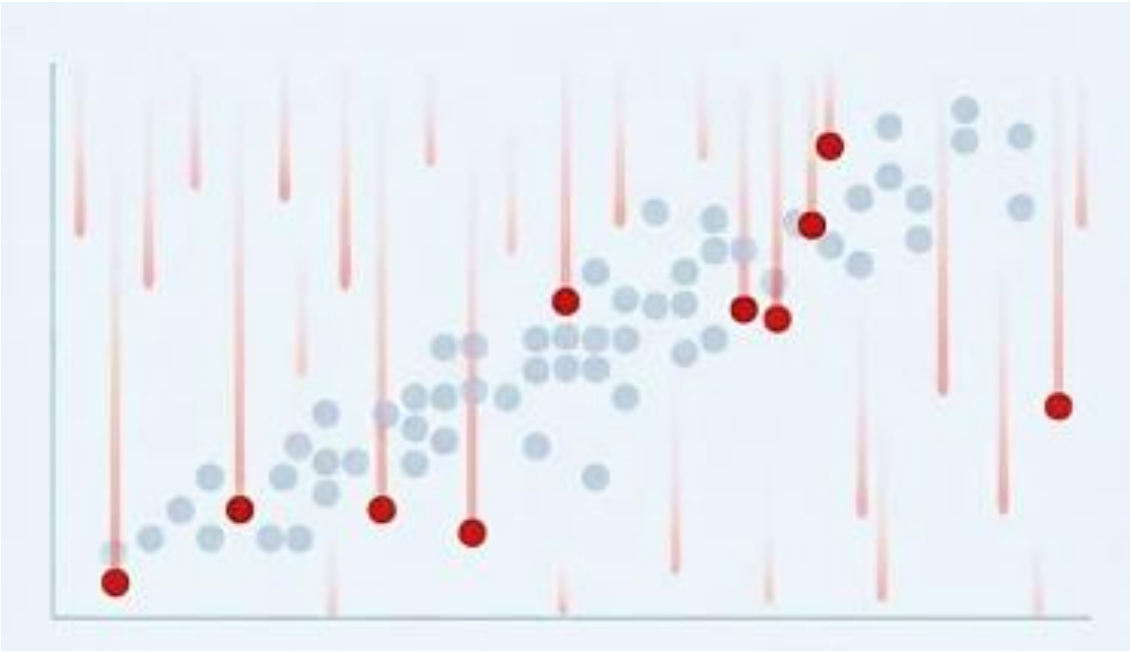
شکل شماره ۲- نمایش شماتیک الگوهای رایج بروز داده‌های گم‌شده در مطالعات پژوهشی؛ شامل گمشدگی تک‌متغیره (متمرکز در یک متغیر)، گمشدگی چندمتغیره (در چند متغیر به‌صورت هم‌زمان)، الگوی یکنواخت یا مونوتون (گمشدگی سیستماتیک در طول زمان، به‌ویژه در مطالعات طولی) و الگوی غیر یکنواخت یا نامنظم (گمشدگی پراکنده و غیرقابل پیش‌بینی در اندازه‌گیری‌های تکرارشونده)

روبین در سال ۱۹۷۶ مکانیسم ایجاد داده‌های گم‌شده را در سه طبقه MCAR، MAR و MNAR طبقه‌بندی کرد (۵). این طبقه‌بندی مبنای بسیاری از روش‌های مدرن تحلیل داده‌های گم‌شده است. در این مقاله تلاش شده است الگوهای کلاسیک گمشدگی شامل MCAR، MARو MNAR نه تنها از نظر مفهومی، بلکه با ارائه مثال عملی و پیامدهای آن‌ها بر تحلیل‌های آماری توضیح داده شوند.

گم شدن کاملا تصادفی (MCAR)

گم شدن داده‌ها به صورت کاملا تصادفی زمانی رخ می‌دهد که احتمال گم شدن هر مشاهده مستقل از مقدار همان متغیر و سایر متغیرهای موجود در داده‌ها باشد. به بیان دیگر در MCAR، فرآیند گم شدن داده‌ها هیچ ارتباط سیستماتیکی با ویژگی‌های مشاهده شده یا مشاهده نشده ندارد (شکل ۳). این فرضیه بسیار قوی است ولی در عمل کمتر اتفاق می‌افتد؛ مگر در شرایطی که داده‌ها به صورت تصادفی یا به دلایل غیر مرتبط، مانند خطاهای فنی در جمع‌آوری داده‌ها، حذف شده باشند (۱۱). یکی از مزایای بارز فرض MCAR در تحلیل آماری، حفظ بی‌طرفی و تخمین بدون سوگیری و با قدرت آماری قابل قبول می‌باشد، با این حال کاهش حجم نمونه می‌تواند قدرت آزمون را کاهش دهد و منجر به برآوردهایی با دقت کمتر شود. به‌طورکلی MCAR مطلوب‌ترین نوع گمشدگی از دیدگاه تحلیل آماری است؛ اما در مطالعات واقعی کمتر مشاهده می‌شود. برای روشن‌تر شدن مفهوم MCAR، می‌توان به کارآزمایی نارسایی قلبی استرالیا و نیوزلند اشاره کرد: این کارآزمایی تصادفی و دوسوکور، کنترل شده با دارونما از یک بتا بلوکر گشادکننده عروق (کارودیلول) در بیماران بالینی پایدار با نارسایی احتقانی قلب و سابقه بیماری عروق کرونر قلب استفاده کرد، مدت زمان پیگیری ۱۸ ماه و هدف اصلی مطالعه تعیین اثرات درمان بر ظرفیت ورزش، عملکرد بطن چپ و اندازه بطن چپ بود (۱۲). در این

کارآزمایی اگر برخی بیماران به صورت کاملاً تصادفی و بدون ارتباط با وضعیت بیماری یا پاسخ درمانی، جلسات پیگیری را از دست داده باشند (برای مثال به دلیل فراموشی)، گمشدگی از نوع MCAR محسوب می‌شود.

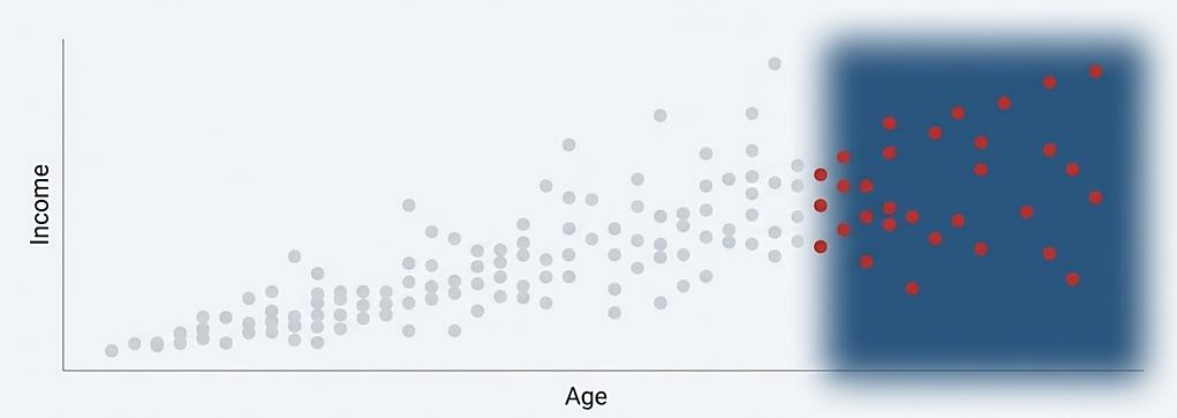


شکل شماره ۳- نمایش مکانیسم گم‌شدگی کاملاً تصادفی (MCAR) که در آن احتمال گم‌شدن داده‌ها مستقل از مقادیر مشاهده شده و مشاهده نشده متغیرها است و گم‌شدگی بدون الگوی سیستماتیک و به‌صورت تصادفی در کل مجموعه داده رخ می‌دهد

گم شدن به صورت تصادفی (MAR)

گم‌شدگی تصادفی زمانی رخ می‌دهد که احتمال گم‌شدن یک مشاهده به متغیرهای مشاهده شده وابسته باشد، اما مستقیماً با مقدار واقعی متغیر گم‌شده ارتباط نداشته باشد. به عنوان نمونه آلیسون (۲۰۰۱) توضیح می‌دهد که در داده‌های مربوط به درآمد، میزان پاسخ ندادن ممکن است به وضعیت تأهل وابسته باشد؛ ممکن است افراد مجرد کمتر از زوج‌های متأهل درآمد خود را گزارش کنند، در این حالت، عدم گزارش درآمد به خود مقدار درآمد وابسته نیست، بلکه ارتباط آن با وضعیت تأهل است (شکل ٤). این نوع مکانیسم گم شدن داده‌ها معمولاً قابل چشم‌پوشی[1] تلقی می‌شود، زیرا با به‌کارگیری روش‌های مناسب (خصوصاً مدل‌هایی که متغیرهای پیش‌بینی‌کننده مرتبط را در خود دارند) می‌توان تحلیل را بدون ایجاد سوگیری معنادار انجام داد. با این حال، هرچه که پیش‌بینی‌کننده‌های مرتبط‌تر و قوی‌تری مانند عوامل خطر یا مواجهه‌هایی که رابطه قوی با داده‌های گم‌شده دارند، در مدل‌های آماری مورد استفاده قرار بگیرند، احتمال برقراری فرض MAR افزایش می یابد و امکان مدیریت معتبر داده‌های گم‌شده فراهم می‌شود. برای روشن‌تر شدن کاربرد MAR در محیط بالینی، می‌توان به کارآزمایی نارسایی قلبی استرالیا و نیوزلند اشاره کرد: در این مطالعه، اگر احتمال غیبت بیمار از جلسات پیگیری به مقادیر قبلاً مشاهده شده (مانند میزان بهبود عملکرد بطن چپ یا ظرفیت ورزش در ویزیت‌های قبلی) وابسته باشد، مثلاً بیمارانی که پاسخ درمانی اولیه ضعیف‌تری داشته‌اند با احتمال بیشتری در جلسات بعد حاضر نشوند، آنگاه گم-شدگی از نوع MAR تلقی می‌شود.

[1] Ignorable

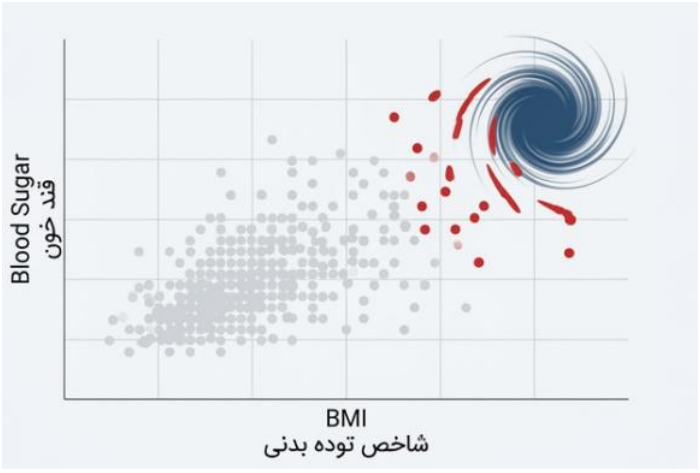


شکل شماره ٤- نمایش مکانیسم گمشدگی تصادفی (MAR) که در آن احتمال گم‌شدن داده‌ها به متغیرهای مشاهده شده وابسته است، اما به مقدار واقعی متغیر گم‌شده وابستگی مستقیم ندارد و می‌توان آن را با استفاده از اطلاعات موجود در داده‌ها مدل‌سازی کرد

گم شدن به صورت غیر تصادفی (MNAR)

در صورتی که الگوی داده‌های گم‌شده با فرضیات مربوط به داده‌های گم‌شده به صورت MCAR یا MAR مطابقت نداشته باشند، در دسته‌بندی گم‌شده‌های غیر تصادفی قرار می‌گیرند. در این حالت، احتمال عدم وجود یک داده مستقیماً به خود داده‌های گم‌شده وابسته است و نمی‌توان آن را صرفا براساس سایر متغیرهای مشاهده شده پیش‌بینی کرد (شکل ۵). این سازوکار گم شدگی به عنوان غیر قابل چشم پوشی[1] شناخته می‌شود زیرا در فرایند مدل‌سازی، نمی‌توان آن را نادیده گرفت. به عنوان مثال در یک بررسی ارتباط دیابت با وزن بدن، ممکن است افراد دارای اضافه وزن، تمایلی به گزارش وزن خود نداشته باشند؛ در این حالت داده‌های مربوط به وزن این افراد، به صورت غیرتصادفی و وابسته به ویژگی‌های فردی از دست می‌روند. برای روشن‌تر شدن این مفهوم در یک محیط بالینی پیچیده‌تر، می‌توان به کارآزمایی نارسایی قلبی استرالیا و نیوزلند اشاره کرد. در این مطالعه، اگر بیمارانی که درمان فعال (کارودیلول) دریافت می‌کنند، به دلیل بروز عوارض جانبی پنهان یا وخامت وضعیت واقعی عملکرد بطن که محققان آن را اندازه‌گیری نکرده‌اند، با احتمال بیشتری جلسات پیگیری را ترک کنند، گم‌شدگی از نوع MNAR محسوب می‌شود. در این شرایط، احتمال عدم حضور در ویزیت وابسته به مقدار واقعی و مشاهده نشده پیامد (مثلاً کاهش عملکرد بطن چپ) است و نه مقادیر مشاهده شده در جلسات قبلی.

[1] Non-ignorable



شکل شماره ۵- نمایش مکانیسم گم‌شدگی غیرتصادفی (MNAR) که در آن احتمال گم‌شدن داده‌ها مستقیماً به مقدار واقعی و مشاهده نشده متغیر گم‌شده وابسته است و این نوع گم‌شدگی می‌تواند منجر به سوگیری قابل‌توجه در تحلیل‌های آماری شود

تشخیص مکانیسم‌های گم شدگی داده‌ها

تحلیل آماری داده‌های گم‌شده مستلزم اتخاذ فرضیه‌های مشخصی درباره مکانیسم‌های گم‌شدگی داده‌ها است. ارزیابی اعتبار این مفروضات پیش از تحلیل، اهمیت به‌سزایی دارد؛ زیرا صحت و قابلیت اعتماد نتایج، منوط به اعتبار فرضیات است. آزمون فرضیه داده‌های گم‌شده به صورت MAR به دلیل نیاز به اطلاعات غیرقابل مشاهده درباره داده‌های گم‌شده، با محدودیت‌های عملی جدی مواجه است. در عمل، تحلیل‌ها اغلب بر پایه فرض MCAR انجام می‌شود، که تنها مکانیسم قابل آزمون آماری است. لیتل آزمون معتبری را برای ارزیابی هم‌زمان فرض MCAR درکل مجموعه ارائه کرده است که در قالب بسته نرم‌افزاری MCARTEST در برنامه SPSS قابل اجرا می‌باشد (۱۳). روش دیگر ساخت شاخص‌های گمشدگی است: برای هر متغیر، یک متغیر دوحالته ایجاد می‌شود (۱ = داده گم‌شده، ۰ = مشاهده شده) سپس آزمون‌های t و آزمون مجذور کای (Chi-square) بین این شاخص و سایر متغیرهای مجموعه داده اجرا می‌شود تا رابطه بین وضعیت گم‌شده بودن و مقادیر سایر متغیرها مورد بررسی قرار گیرد (۱٤). عدم معناداری، حاکی از عدم رد فرض MCAR است (هرچند امکان وجود MAR یا NMAR باقی می‌ماند). به عنوان نمونه، اگر زنان نسبت به مردان، سن واقعی خود را کمتر گزارش دهند؛ آزمون مجذور کای ممکن است تفاوت معناداری در درصد داده‌های گم‌شده سن بر اساس جنسیت آشکار سازد. نرم افزار SPSS قابلیت اجرای خودکار این تحلیل‌ها را دارا است.

روش‌های برخورد با داده‌های گم‌شده

یکی از چالش‌های اساسی در مواجهه با داده‌های گم‌شده، ایجاد سوگیری انتخابی است که می‌تواند بر برآوردها و استنتاج‌های آماری تاثیر قابل توجهی بگذارد. هدف اصلی روش‌های اصولی مدیریت داده‌های گم‌شده، کاهش یا حذف این سوگیری‌ها و تضمین برآوردهای معتبر اثرات واقعی است (۱۵). بهترین استراتژی، پیشگیری از بروز داده‌های گم‌شده از طریق طراحی دقیق مطالعه و جمع‌آوری صحیح داده‌ها می‌باشد (۱٦, ۱۷). در واقع، پیشگیری بسیار مؤثرتر و مطمئن‌تر از روش‌های اصلاحی پس از وقوع گم‌شدگی است. هرچند داده‌های گم‌شده در بسیاری از پژوهش‌ها (به‌ویژه مطالعات طولی) پدیده‌ای طبیعی و تا حدی اجتناب‌ناپذیر به شمار می‌آیند، تحلیل‌گران باید از رویکردهایی بهره ببرند که در برابر نقض خفیف تا متوسط فرضیات مقاوم باشند و میزان سوگیری در استنباط نسبت به جامعه هدف را به حداقل

برسانند. با این حال به‌کارگیری روش‌های پیشرفته همیشه ممکن یا عملی نیست، از این‌رو توسعه و معرفی رویکردهای جایگزین و قابل‌اجرا برای مدیریت داده‌های گم‌شده اهمیت روزافزونی پیدا کرده است (۹). در شرایطی که میزان داده‌های گم‌شده در مطالعه کم باشد، معمولاً تفاوت‌های حاصل از کاربرد روش‌های مختلف مدیریت داده‌های گم‌شده، تأثیر قابل توجهی بر نتایج نخواهد داشت (۱۸). پس از معرفی و توضیح روش‌های مختلف مدیریت داده‌های گم‌شده، در این بخش به منظور تبیین عملی اهمیت داده‌های گم‌شده و تأثیر روش‌های متفاوت مدیریت آن‌ها بر نتایج تحلیل‌های آماری، یک مثال کوچک از مطالعه‌ای فرضی در بیماران دیابتی ارائه می‌شود (جدول ۱). این مثال با ساختار متداول داده‌های موجود در سامانه‌های سلامت ایران سازگار است و وضعیت رایج گم‌شدگی در مقادیر HbA1c را که یکی از چالش‌های روزمره در پژوهش‌های دیابت است، بازنمایی می‌کند. در این مثال، ۱۰ بیمار طی یک دوره ٦ ماهه پیگیری شدند، BMI به‌عنوان متغیر پیش‌بینی‌کننده و HbA1c به‌عنوان متغیر پیامد در نظر گرفته شد. برای ۳ نفر از شرکت‌کنندگان مقدار HbA1c ثبت نشده بود.

جدول شماره ۱- داده‌های مثال بالینی از یک مطالعه کوهورت فرضی در بیماران دیابتی طی دوره پیگیری شش‌ماهه، شامل شاخص توده بدنی (BMI) به‌عنوان متغیر پیش‌بینی‌کننده و HbA1c به‌عنوان متغیر پیامد، با وجود مقادیر گم‌شده در HbA1c برای بخشی از بیماران

ID	BMI	HbA1c
۱	۲۵/۱	٦/۲
۲	۳۱/٤	۷/۸
۳	۲۸/۷	...
٤	۳۳/۹	۸/۱
۵	۲۷/۵	٦/۹
٦	۲۹/۳	...
۷	۲٦/٤	٦/۵
۸	۳۲/۱	۷/۹
۹	۳۰/۷	...
۱۰	۲٤/۹	۵/۸

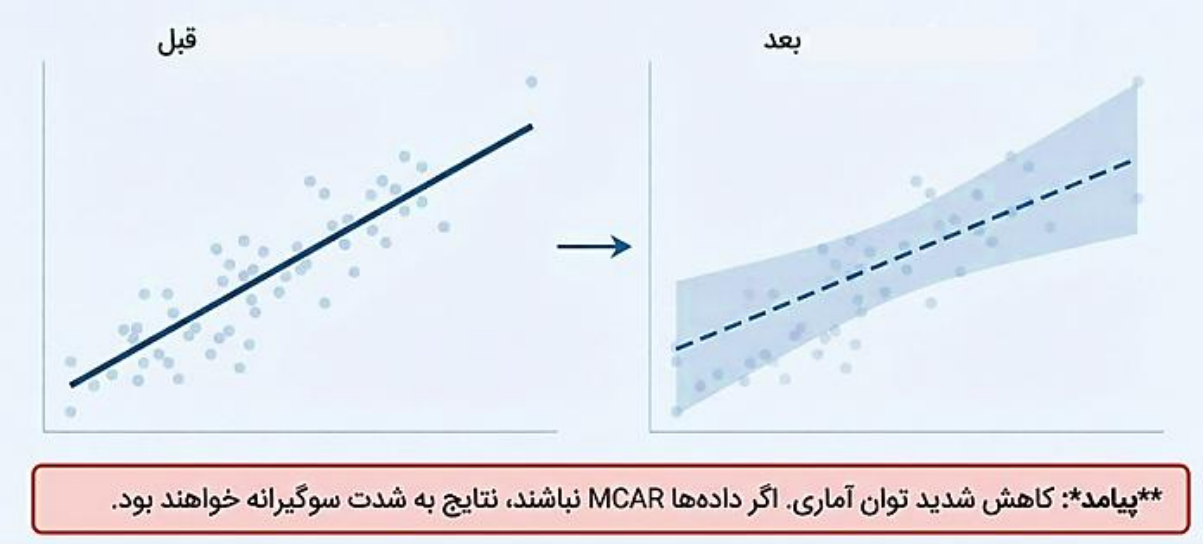
در ادامه، نحوه مدیریت داده‌های گم‌شده با استفاده از چهار روش مختلف در نرم افزار استتا توضیح داده شده است. این روش‌ها شامل حذف مشاهدات، جای‌گذاری مقادیر گم‌شده با مقادیر میانگین، استفاده از مدل‌های رگرسیونی و همچنین تکنیک‌های پیشرفته‌تر مانند جای‌گذاری چندگانه هستند.

تحلیل موارد کامل[1](حذف لیستی[2])

تحلیل موارد کامل یکی از ساده‌ترین و در عین حال رایج‌ترین رویکردها برای مدیریت داده‌های گم‌شده است که در بسیاری از نرم افزارهای آماری نیز به عنوان روش پیش‌فرض به‌کار می‌رود. در این روش، تنها مواردی وارد

[1] Complete case analysis
[2] Listwise deletion

تحلیل می‌شوند که تمام متغیرهای مورد مطالعه به طورکامل مشاهده شده‌اند، و هر مشاهده‌ای که دارای مقدار گم‌شده باشد از تجزیه و تحلیل کنار گذاشته می‌شود. اگر گم‌شدگی داده‌ها از نوع MCAR باشد، موارد کامل می‌توانند نمونه‌ای تصادفی از جمعیت اولیه محسوب شوند و بنابراین، برآوردها بدون سوگیری خواهند بود. با این حال در صورت وجود NMAR یا MAR حذف لیستی منجر به برآوردهای دارای سوگیری خواهد شد، زیرا موارد کامل نمی‌توانند نمایانگر نمونه اصلی باشند (۱۹). علاوه بر مسئله سوگیری، استفاده از تحلیل موارد کامل معمولاً باعث کاهش چشمگیر حجم نمونه می‌شود؛ موضوعی که در مطالعاتی مانند متاآنالیزها یا پژوهش‌هایی با حجم نمونه محدود تأثیر بیشتری دارد. این رویکرد اطلاعات ارزشمند موجود در داده‌های گم‌شده را نادیده می‌گیرد که می‌تواند اعتبار و دقت نتایج را کاهش دهد (۶, ۲۰) (شکل ۶).



شکل شماره ۶- نمایش شماتیک روش حذف کامل موارد (Complete Case Analysis) که در آن تنها مشاهدات دارای داده‌های کامل وارد تحلیل می‌شوند و تمام موارد دارای حداقل یک مقدار گم‌شده از فرایند تحلیل آماری حذف می‌گردند، امری که می‌تواند منجر به کاهش حجم نمونه شود

همان‌طور که پیش‌تر اشاره شد، در مطالعه فرضی که ۱۰ بیمار دیابتی طی یک دوره ۶ ماهه پیگیری شدند، مقدار HbA1c برای ۳ بیمار گم شده است. در تحلیل موارد کامل، تنها مشاهداتی که مقدار HbA1c آن‌ها ثبت شده است وارد مدل خواهند شد و تحلیل رگرسیونی بر اساس داده‌های کامل انجام می‌شود.

reg hba1c bmi if hba1c<.

نتایج نشان می دهد که بین BMI و HbA1c رابطه‌ای مثبت و معنی‌دار وجود دارد (p<۰/۰۰۱، β=۰/۲۲۵). این مدل ۹۸ درصد از تغییرات HbA1c را تبیین می‌کند (R^2 =۰/۹۸۳). اگرچه این روش از نظر اجرایی ساده است، اما تحلیل مبتنی بر تنها ۶ مشاهده باعث کاهش توان آماری و افزایش احتمال بیش‌برازش می‌شود. علاوه‌براین، این روش فقط در صورتی برآوردهای بدون سوگیری ارائه می‌کند که داده‌ها به‌طور MCAR گم شده باشند؛ در غیر این‌صورت تخمین‌ها قابل اعتماد نیستند.

تجزیه و تحلیل موارد موجود[1] (حذف زوجی[2])

تحلیل موارد موجود که آنالیز زوجی نیز نامیده می‌شود، روشی است که برای برآورد شاخص‌های آماری، به‌ویژه همبستگی، از بیشترین مقدار داده‌های در دسترس استفاده می‌کند. در این رویکرد، اندازه نمونه برای هر محاسبه متفاوت است؛ زیرا برای برآورد هر همبستگی تنها

[1] Available Case Analysis
[2] Pairwise Deletion

مواردی به کار می‌روند که هر دو متغیر موردنظر مقدار غیرگم‌شده دارند. برای مثال، اگر سه متغیر A، B و C به ترتیب ۹۰، ۷۰ و ۶۰ درصد داده موجود داشته باشند، همبستگی A و B با نمونه بزرگ‌تری نسبت به همبستگی A و C محاسبه می‌شود و همبستگی B و C نیز از کوچک‌ترین زیرنمونه به دست می‌آید. همین تغییر اندازه نمونه میان زیرمجموعه‌های مختلف می‌تواند به ناسازگاری و دشواری در تفسیر نتایج منجر شود؛ به‌ویژه زمانی که میزان گم‌شدگی هر متغیر متفاوت باشد. در شرایطی که داده‌ها بر اساس مکانیزم MCAR حذف شده باشند، زیرنمونه‌ها نماینده داده‌های اصلی هستند و تحلیل موارد موجود برآوردهای بی‌طرفانه ارائه می‌دهد؛ با این حال، تحت مکانیزم MAR، زیرنمونه‌ها نماینده جمعیت اصلی نیستند و نتایج می‌توانند سوگیرانه باشند. علاوه بر این، به‌دلیل اینکه هر عنصر ماتریس واریانس-کوواریانس از زیرنمونه متفاوتی به دست می‌آید، این ماتریس ممکن است غیرقابل معکوس شدن شود و محاسبه خطاهای استاندارد نیز، همان‌طور که پل آلیسون اشاره کرده است با چالش و احتمال خطا همراه باشد. هرچند تحلیل موارد کامل به‌عنوان رویکردی ساده و مبتنی بر داده‌های در دسترس مطرح است، اما در صورت وجود داده‌های گمشده با الگوی غیرتصادفی، می‌تواند منجر به سوگیری در برآوردها و کاهش اعتبار نتایج شود. از این‌رو، استفاده از این روش باید با احتیاط و در کنار تحلیل‌های تکمیلی به‌منظور ارزیابی پایداری نتایج صورت گیرد.

جای‌گذاری[1]

یکی از روش‌های پرکاربرد برای مدیریت داده‌های گم-شده روش جای‌گذاری است، که در آن به جای حذف مقادیر گم‌شده، این مقادیر با مقادیر تخمینی جایگزین می‌شوند. هدف این رویکرد حفظ کامل مجموعه داده‌ها و استفاده کامل از اطلاعات موجود است، بدین معنا که با جایگزینی داده‌های گم‌شده با مقادیر احتمالی، تمامی موارد در تحلیل باقی می‌مانند و از کاهش حجم نمونه جلوگیری می‌شود. پس از انجام جای‌گذاری، مجموعه داده‌ها با استفاده از روش‌های استاندارد برای داده‌های کامل، مورد تحلیل قرار می‌گیرند. روش جای‌گذاری به‌طور کلی به دو دسته تقسیم می‌شود: جای‌گذاری تک مقداری و جای‌گذاری چندگانه MI[2]. روش‌های جای-گذاری تک مقداری، مانند استفاده از میانگین، رگرسیون خطی، یا استفاده از آخرین مقدار مشاهده شده برای داده‌های طولی، به طور معمول توصیه نمی‌شوند، زیرا با کاهش مصنوعی واریانس، منجر به کاهش کاذب خطاهای استاندارد و در نتیجه ایجاد اعتماد بیش از حد به برآوردها می‌شوند (۵). رویکرد جای‌گذاری چندگانه روش توصیه شده و استانداردتر است، که شامل سه مرحله اصلی می‌باشد: تولید چندین مجموعه داده با مقادیر جای‌گذاری شده، تحلیل جداگانه هر مجموعه داده و ادغام نتایج طبق قواعد روبین برای به‌دست آوردن برآوردهای نهایی (۱۷). در ادامه، رایج‌ترین روش‌های جای‌گذاری را به تفصیل بررسی می‌کنیم.

انواع جایگذاری تک مقداری[3]

جای‌گذاری میانگین، میانه یا مد:

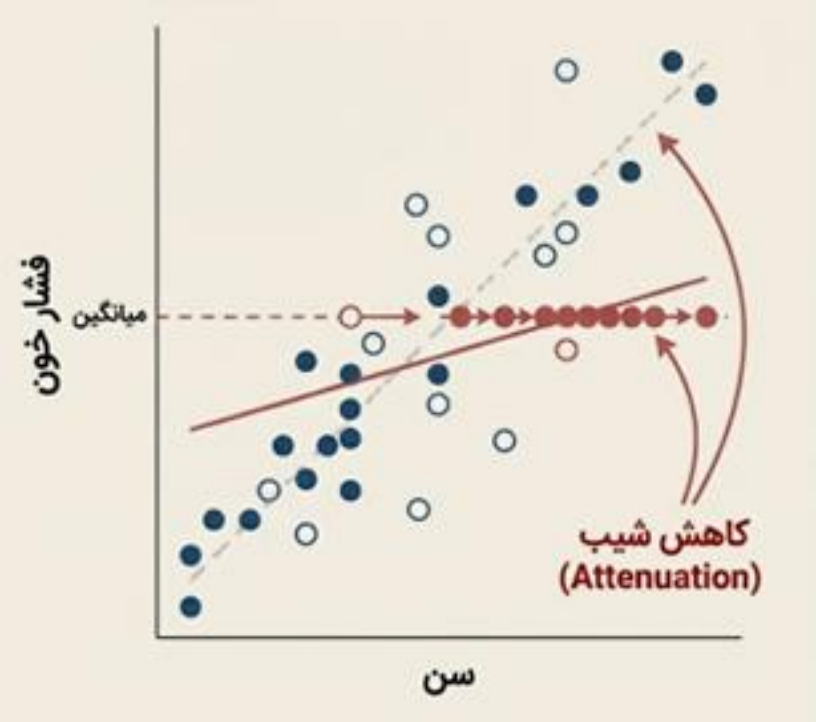
این روش یکی از ساده‌ترین روش‌های جای‌گذاری تک مقداری است که در آن مقدار میانگین مشاهده شده برای یک متغیر به عنوان جایگزین مقادیر گم‌شده قرار می‌گیرد (شکل ۷). این روش ضمن فرض MCAR بودن داده‌ها، ساختار طبیعی پراکندگی داده‌ها را نادیده می‌گیرد و به همین دلیل در اغلب کاربردهای تحلیلی توصیه نمی‌شود، زیرا باعث تجمع مصنوعی نقاط اطراف میانگین متغیر وابسته، کاهش فواصل اطمینان و برآوردهای کمتر واریانس شده در نتیجه منجر به سوگیری در نتایج تحلیل‌های آماری می‌شود (۱۲, ۱۵). جای‌گذاری با میانه در داده‌های چول یا زمانی که مقادیر پرت وجود دارد مفید است، زیرا میانه نسبت به مقادیر افراطی مقاوم‌تر است. این روش ساختار تنوع داده را کاهش می‌دهد و

[1] Imputation

[2] Multiple imputation

[3] Single imputation

می‌تواند روابط آماری واقعی را تضعیف کند. در جایگذاری مد، مقادیر گم‌شده به‌وسیله پرتکرارترین مقدار (مد) هر متغیر پر می‌شوند. این روش به‌ویژه برای متغیرهای طبقه‌ای مناسب است و اجرای بسیار ساده‌ای دارد. با این حال، می‌تواند باعث افزایش مصنوعی فراوانی یک دسته خاص شود و در نتیجه توزیع واقعی داده را مخدوش کند.



شکل شماره ۷- نمایش روش جای‌گذاری میانگین که در آن مقادیر گم‌شده هر متغیر با میانگین مقادیر مشاهده شده همان متغیر جایگزین می‌شوند؛ رویکردی ساده که اگرچه حجم نمونه را حفظ می‌کند، اما باعث کاهش واریانس و تضعیف روابط آماری واقعی می‌شود

.

فرض کنید در مطالعه قبلی که ۱۰ بیمار دیابتی طی ٦ ماه پیگیری شدند و مقدار HbA1c برای ۳ بیمار گم شده است، از روش جای‌گذاری با میانگین برای مقادیر گم-شده استفاده می‌شود. برای انجام این کار در نرم افزار استتا، ابتدا میانگین مقدار HbA1c محاسبه شده و سپس مقادیر گم‌شده با آن میانگین جایگزین می‌شود:

```
summ hba1c
replace hba1c = r(mean) if hba1c== .
reg hba1c bmi
```

پس از جای‌گذاری مقادیر گم‌شده با میانگین، تعداد مشاهدات به ۱۰ افزایش می‌یابد و رابطه همچنان معنی دار باقی می‌ماند: (β =۰/۲۳۵ ،p <۰/۰۰۱ ،R^2 =۰/۹۲۵)

اگرچه این روش ممکن است در ظاهر نتایج قابل قبولی ارائه دهد، اما به دلیل مشکلات و محدودیت‌های مطرح‌شده، استفاده از آن توصیه نمی‌شود.

رویکرد آخرین مشاهده حمل شده به جلو (LOCF[1]) و مشاهده پایه حمل شده به‌جلو (BOCF[2])

رویکرد آخرین مشاهدات حمل شده به جلو یکی از روش‌های سنتی رسیدگی به داده‌های گم‌شده درکارآزمایی‌های بالینی است که در آن آخرین مقدار مشاهده شده برای هر فرد به عنوان جایگزین مقادیر گم-شده به‌کار می‌رود. اگرچه این روش در گذشته کاربرد

[1] Last Observation Carried Forward
[2] Baseline Observation Carried Forward

گسترده‌ای داشته است، اما در سال‌های اخیر به دلیل احتمال ایجاد سوگیری قابل توجه در برآورد اثر درمان مورد انتقاد قرار گرفته است. هرچند LOCF نسبت به حذف فهرستی از نظر حفظ حجم نمونه روش مناسب‌تری است، اما به‌دلیل تحریف روند واقعی پاسخ و کاهش توان تحلیل با محدودیت‌های جدی همراه است. رویکرد مشاهده پایه به جلو BOCF که اخیراً در ادبیات بالینی توجه بیشتری به آن شده است، منطق مشابهی با LOCF دارد، با این تفاوت که در این روش مقادیر خط پایه برای جایگزینی داده‌های گم‌شده استفاده می‌شود. قابل ذکر است که استفاده از BOCF زمانی معنا دارد که فرد همچنان از نظر وضعیت پاسخ تحت ارزیابی قرار گیرد؛ با این حال، در صورت خروج از مطالعه (صرف‌نظر از زمان یا دلیل آن) مقدار خط پایه به‌عنوان پاسخ نهایی فرد در نظر گرفته می‌شود (۲۱, ۲۲).

رگرسیون خطی[1] در تحلیل داده های گم‌شده

یکی از رویکردهای رایج برای تحلیل داده‌های گم‌شده جای‌گذاری مبتنی بر رگرسیون خطی است که در آن یک معادله رگرسیون با استفاده از داده‌های کامل برآورد می‌شود و سپس مقادیر پیش‌بینی شده برای جانشینی مقادیر گم‌شده متغیر نتیجه به‌کار می‌روند. در این روش، متغیرهای پیش‌بینی کننده مرتبط در مدل وارد می‌شوند تا بهترین برآورد ممکن از داده گم‌شده ایجاد شود. این رویکرد از نظر مفهومی مشابه جای‌گذاری با میانگین است، اما با استفاده از اطلاعات چندین پیش‌بینی کننده و افزودن مولفه‌های تصادفی در نسخه‌های پیشرفته‌تر آن، می‌تواند بخشی از عدم‌قطعیت موجود در داده‌ها را بازتاب دهد و از بروز سوگیری ناشی از ثابت بودن مقدار جایگزین جلوگیری کند. این روش تحت فرض MAR معتبر است. با این حال، در مواردی که ساختار داده‌ها و الگوهای گم‌شدگی پیچیده‌تر باشد، تعیین وزن‌ها و محاسبه خطاهای استاندارد می‌تواند بسیار دشوار و پیچیده شود (۱۲). در نتیجه، بهره‌گیری از این روش نیازمند تحلیل دقیق فرضیات و ملاحظات مربوط به ساختار داده‌ها است.

در مثال قبلی، همان‌طور که ذکر شد، ۳ بیمار از مطالعه دیابتی دارای داده‌های گم‌شده از HbA1c هستند. برای استفاده از روش رگرسیون خطی به‌عنوان تک‌تخصیص، ابتدا یک مدل رگرسیونی با استفاده از داده‌های کامل برازش می‌شود و سپس مقادیر پیش‌بینی‌شده برای این سه بیمار به‌عنوان جایگزین مقادیر گم‌شده استفاده می‌شوند:

```
reg hba1c bmi if hba1c <.
predict hba1c_hat if hba1c == .
replace hba1c = hba1c_hat if hba1c == .
reg hba1c bmi
```

پس از انجام این مراحل، نتیجه نهایی نشان داد که اثر BMI همچنان قوی و معنی‌دار باقی می‌ماند (β=۰/۲۴۵ ،p <۰/۰۰۱ ،R^2 = ۰/۹۶۷).

این روش نسبت به جای‌گذاری با میانگین، عملکرد بهتری دارد و قادر است تا از اطلاعات چندین پیش‌بینی‌کننده برای بهبود برآورد مقادیر گم‌شده استفاده کند. با این حال، همانند روش جای‌گذاری میانگین، تک‌تخصیص رگرسیونی عدم‌قطعیت ناشی از گمشدگی داده‌ها را به‌طور کامل لحاظ نمی‌کند و معمولاً به برآوردهای بیش‌اعتمادانه منجر می‌شود. به همین دلیل، این روش نیز نمی‌تواند به‌طور کامل جایگزین روش‌های پیشرفته‌تر مانند جایگذاری چندگانه شود که می‌توانند به‌طور دقیق‌تر عدم‌قطعیت و واریانس را بازتاب دهند.

جای‌گذاری هم‌گروه[2]

روش جای‌گذاری هم‌گروه، شامل جای‌گذاری مقادیرگم-شده با مقادیر مربوط به پاسخ‌دهندگان دیگر با ویژگی‌های همسان است. این فرآیند با شناسایی و گروه‌بندی شرکت‌کنندگان بر اساس شباهت در داده‌های مشاهده شده، اقدام به برآورد مقادیر گم-شده می‌کند. از ویژگی‌های مهم این روش، سادگی نسبی است، که در آن نسبت به روش‌های جای‌گذاری میانگین یا استفاده کامل از داده‌های مشاهده شده، نتایج دچار سوگیری کمتری

[1] Linear regression

[2] Hot Deck Imputation

می‌شود. فرض این رویکرد، MAR بودن داده‌ها است. یکی از محدودیت‌های آن، نیاز به استفاده از الگوریتم‌های تطابق پیچیده، در مواردی که تطابق پاسخ‌دهندگان بر اساس ویژگی‌ها دشوار است، می باشد (۱۲).

جای‌گذاری بر اساس منبع ثابت[1]

جای‌گذاری بر اساس منبع ثابت، روشی است که در آن مقادیر گم‌شده با استفاده از اطلاعات خارجی معتبر (مانند داده‌های مطالعات قبلی، بانک‌های اطلاعاتی استاندارد یا دانش قبلی موجود) جایگزین می‌شوند. این رویکرد در مواردی کاربرد دارد که داده‌های خارجی معتبر در دسترس باشد و بتوانند نماینده‌ای مناسب برای متغیرهای گم‌شده فراهم کنند. با این حال، یکی از معایب کلیدی آن وابستگی شدید به کیفیت و اعتبار داده‌های خارجی است، به‌گونه‌ای که هرگونه خطا یا ناسازگاری در منابع بیرونی می‌تواند مستقیماً موجب سوگیری و کاهش اعتبار تحلیل‌های بعدی شود (۱۲).

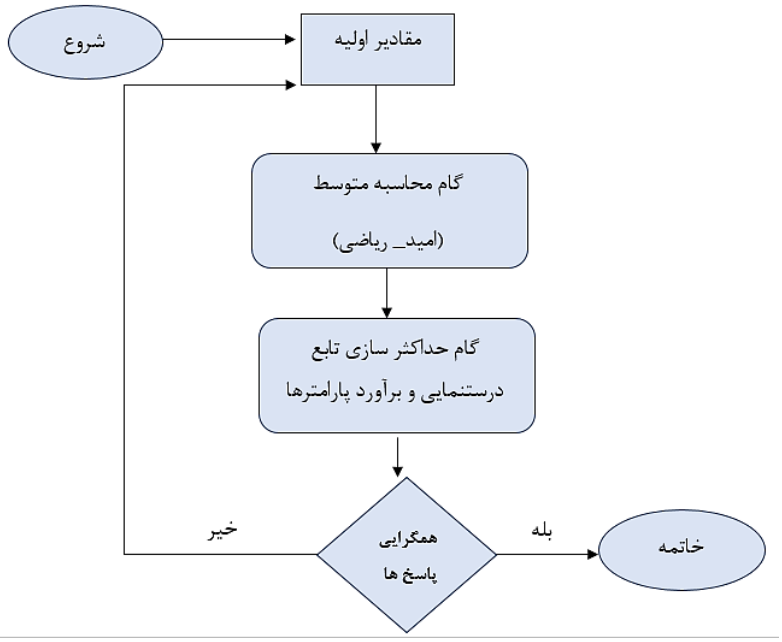
حداکثر درست‌نمایی (MLE[2])

روش حداکثر درست‌نمایی یکی از رویکردهای معتبر برای تحلیل داده‌های گم‌شده است. در این روش فرض می‌شود که داده‌های مشاهده شده از یک توزیع نرمال چند متغیره به‌دست آمده‌اند. سپس پارامترهای مدل (میانگین‌ها، کوواریانس‌ها و سایر ضرایب) با حداکثر تابع درست‌نمایی مبتنی بر داده‌های موجود، برآورد می‌شوند و روابط بین متغیرها بدون نیاز به تکمیل مستقیم داده‌های گم‌شده برآورد می‌گردد. روش MLE به‌ویژه در مدل‌های چندمتغیره و زمانی که میزان گم‌شدگی نسبتاً کم باشد، رویکردی کارآمد، سازگار و بدون سوگیری می‌باشد (۹). با وجود مزایای این روش، محدودیت مهم آن این است که همانند بسیاری از روش‌های مبتنی بر احتمال، تحت مفروضات MAR معتبر است؛ اما برای داده‌های MNAR معتبر نیست، زیرا هیچ الگوی واحد یا قابل‌اجماعی برای مدل‌سازی مکانیسم MNAR وجود ندارد (۱۵). از این‌رو، استفاده از MLE در شرایطی که احتمال گم‌شدگی غیرتصادفی وجود دارد، نیازمند تحلیل حساسیت و بررسی مفروضات جایگزین است.

امید ریاضی–بیشینه سازی (EM[3])

الگوریتم امید ریاضی-بیشینه‌سازی یکی از مهم‌ترین روش‌های مبتنی بر حداکثر درست‌نمایی برای تحلیل داده‌های گم‌شده است که توانایی تولید مجموعه داده‌های جدید را دارد (۲۳). در این روش، پارامترهای مدل با استفاده از یک فرآیند تکراری دو مرحله‌ای برآورد می‌شوند. در مرحله اول یعنی مرحله انتظار داده‌های گم-شده با استفاده از پارامترهای فعلی مدل به‌طور ضمنی با مقادیر پیش‌بینی شده جایگزین می‌شوند و سپس آماره‌هایی مانند میانگین‌ها و واریانس‌ها برآورد می‌شوند. در مرحله بیشینه سازی، پارامترهای جدید با بیشینه سازی تابع درست‌نمایی کامل شده (مبتنی بر داده‌های واقعی و داده‌های جای‌گذاری شده در مرحله انتظار) محاسبه می‌شوند. این دو مرحله به صورت تکراری ادامه پیدا می‌کند تا زمانی که پارامترها به هم‌گرایی برسند (شکل ۸).

[1] Cold deck imputation
[2] Maximum Likelihood Estimation
[3] Expectation-maximization

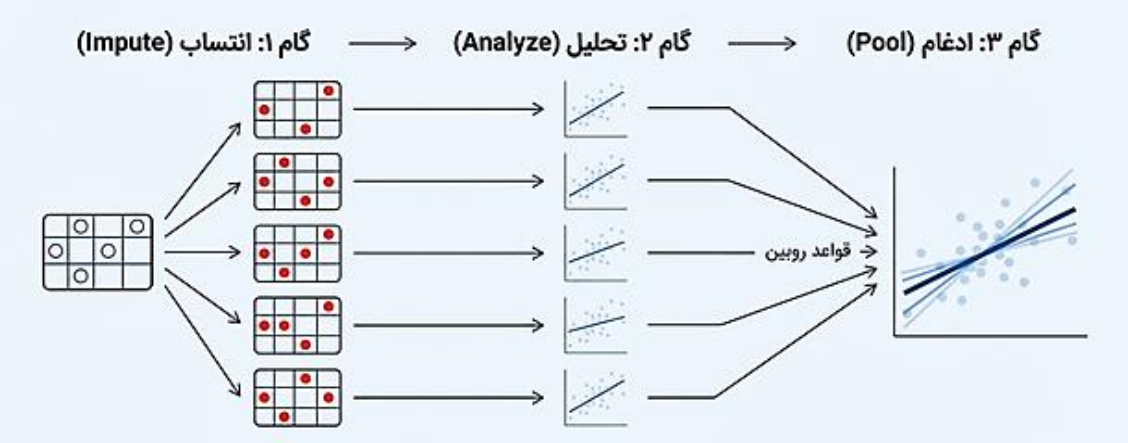


شکل شماره ۸- الگوریتم امید ریاضی_بیشینه سازی

الگوریتم EM در شرایطی که میزان گم‌شدگی کم تا متوسط باشد عملکرد خوبی دارد، اما در صورت گم-شدگی زیاد ممکن است روند هم‌گرایی طولانی شود یا اصلاً رخ ندهد و پیچیدگی محاسباتی افزایش یابد. همچنین ممکن است EM منجر به برآوردهای مغرضانه یا کاهش خطای استاندارد شود که به تفسیر نادرست نتایج منجر می‌شود (۹). علاوه بر این، مانند دیگر روش‌های مبتنی بر احتمال، EM تحت فرض MAR معتبر است و برای شرایط MNAR به‌تنهایی قابل اعتماد نیست. بنابراین، استفاده از EM نیازمند بررسی دقیق الگوی گمشدگی، ارزیابی هم‌گرایی و آگاهی از محدودیت‌های تفسیری آن است.

جای‌گذاری چندگانه (MI)

جای‌گذاری چندگانه یکی از معتبرترین و پرکاربردترین رویکردها برای مدیریت داده‌های گم‌شده است. این فرایند در سه گام اصلی انجام می‌پذیرد (شکل ۹). در گام اول، مقادیر گم‌شده هر متغیر بر اساس اطلاعات موجود در سایر متغیرها و بر اساس یک مدل آماری مناسب پیش‌بینی می‌شود. مقادیر پیش‌بینی شده، که "مقادیر جای‌گذاری" نامیده می‌شوند، جایگزین مقادیر گم‌شده می‌شوند و یک مجموعه داده کامل با عنوان "مجموعه داده جای‌گذاری شده" ایجاد می‌شود. این فرآیند چندین بار تکرار شده و به‌جای یک مجموعه داده کامل، چندین مجموعه داده جای‌گذاری شده تولید می‌گردد که هر کدام بازتابی از عدم‌قطعیت ناشی از گم‌شدگی هستند. در گام دوم، تحلیل‌های آماری استاندارد به طور جداگانه بر روی هر مجموعه انجام می‌پذیرد و در نهایت نتایج حاصل طبق قواعد روبین برای به‌دست آوردن یک برآورد نهایی تلفیق می‌گردند. روش جای‌گذاری چندگانه نه تنها به بازنمایی تغییرپذیری طبیعی موجود در داده‌های گم‌شده کمک می‌کند، بلکه با لحاظ کردن عدم قطعیت ناشی از تخمین مقادیر گم‌شده روش معتبرتری را ارائه می‌دهد. مطالعات شبیه‌سازی و تجربی متعددی نشان داده‌اند که MI قادر است در طیف وسیعی از سناریوهای گم‌شدگی برآوردهای تقریباً بدون سوگیری و قابل اعتمادی را فراهم کند (۲٤). با توجه به محدودیت حجم این مقاله، امکان ارائه توضیحات جامع و گسترده درباره جای‌گذاری چندگانه وجود ندارد؛ از این‌رو تنها به مرور کوتاهی از چند نمونه از پرکاربردترین روش‌های آن بسنده شده است.



شکل شماره ۹- مراحل اصلی جای‌گذاری چندگانه (MI) شامل تولید چندین مجموعه‌داده کامل با جای‌گذاری مقادیر گم-شده بر اساس مدل آماری مناسب، انجام تحلیل جداگانه بر روی هر مجموعه‌داده و ترکیب نتایج نهایی طبق قواعد روبین به‌منظور لحاظ کردن عدم‌قطعیت ناشی از داده‌های گم‌شده

.

تخصیص کاملاً شرطی FCS[1]

تخصیص کاملاً شرطی رویکردی انعطاف‌پذیر برای جای‌گذاری داده‌های گم‌شده است که در آن مقادیر گم-شده به صورت متغیر به متغیر و بر اساس یک مدل شرطی مجزا برای هر متغیر برآورد می‌شوند. در این روش، برای هر متغیر گم‌شده یک مدل پیش‌بینی مناسب، متناسب با نوع متغیر (پیوسته، دودویی، دسته‌ای و ...) تعریف شده و فرآیند جای‌گذاری به‌صورت تکراری میان متغیرها اجرا می‌گردد. از میان روش‌های مبتنی بر FCS، جای‌گذاری چندگانه با استفاده از معادلات زنجیره‌ای (MICE[2]) رایج‌ترین و پرکاربردترین رویکرد است. در MICE، متغیرهای دارای گمشدگی با استفاده از سایر متغیرهای موجود مدل سازی و مقدار دهی می‌شوند و این چرخه تا رسیدن به هم‌گرایی، ادامه می‌یابد. نتیجه این تکرارها تولید مجموعه داده‌های کاملی است که عدم قطعیت ناشی از گمشدگی را منعکس می‌کنند. روش MICEبه دلیل انعطاف‌پذیری، سازگاری با انواع داده‌ها و سهولت پیاده‌سازی در نرم‌افزارهای آماری، به‌عنوان یکی از معتبرترین روش‌های جای‌گذاری چندگانه در مدیریت داده‌های گم‌شده شناخته می‌شود و در بسیاری از مطالعات اپیدمیولوژیک و بالینی مورد استفاده قرار می‌گیرد.

می‌توان به مثال قبلی که در آن ۱۰ بیمار دیابتی طی ٦ ماه پیگیری شدند، اشاره کرد. در این مطالعه، مقدار HbA1c برای ۳ بیمار گم شده بود. در اینجا از روش MICE برای مدیریت داده‌های گم‌شده استفاده شد و مقادیر گم‌شده برای HbA1c با استفاده از مدل‌های رگرسیونی و سایر متغیرهای موجود مانند BMI جایگزین شدند. دستورات استتا به شکل زیر اجرا شد:

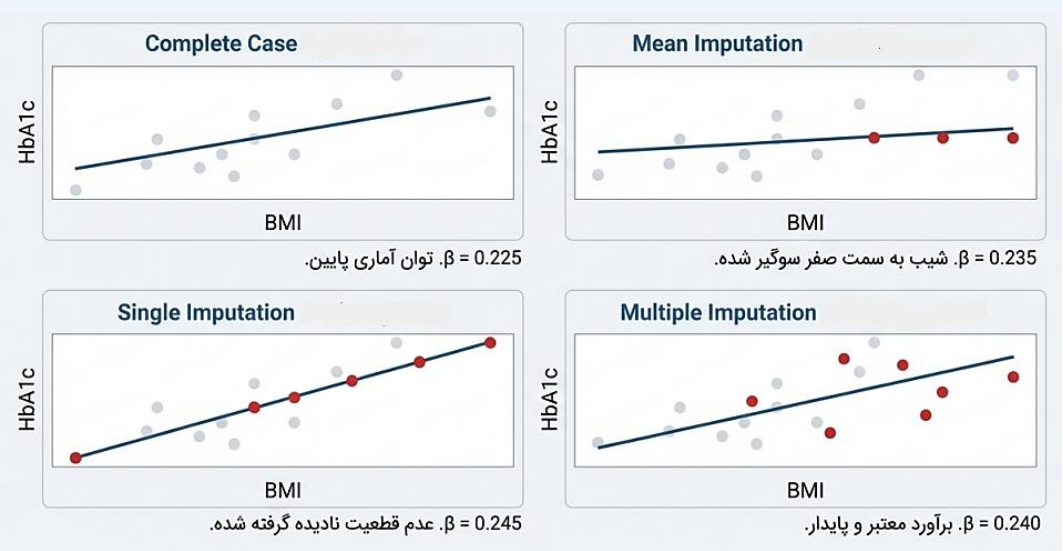
```
mi set wide
mi register imputed hba1c
mi impute chained (regress) hba1c = bmi,
add(5) rseed(۱۲۳٤٥)
mi estimate: regress hba1c bmi
```

پس از اجرای این دستورات، ۵ مجموعه داده جای‌گذاری شده تولید شدند که در آن‌ها مقادیر گم‌شده HbA1c با استفاده از مدل رگرسیونی پیش‌بینی شدند. سپس، مدل رگرسیونی برای بررسی رابطه بین HbA1c و BMI برآورد شد. در این مثال، اثر BMI بر HbA1c در تمامی مجموعه داده‌های جای‌گذاری شده همچنان معنی‌دار باقی ماند (β=۰/۲٤۰ ،p<۰/۰۰۱).

در مقایسه با روش‌هایی مانند حذف موارد، جای‌گذاری با میانگین و رگرسیون خطی، MICE تخمین‌های پایدارتر و معتبرتری ارائه داد. در این مطالعه، MICE توانست

[1] Fully conditional specification
[2] Multiple Imputation by Chained Equations

اطلاعات گم‌شده را به‌طور معتبر جایگزین کند و عدم‌قطعیت موجود در داده‌های گم‌شده را به‌خوبی در نتایج تحلیل‌ها بازتاب دهد. در نتیجه، این روش به‌ویژه در مطالعات پزشکی و اپیدمیولوژیک، از جمله پژوهش‌های مرتبط با دیابت، به‌عنوان بهترین روش برای مدیریت داده‌های گم‌شده شناخته می‌شود. مقایسه بصری نتایج حاصل از چهار رویکرد مدیریت داده‌های گم‌شده در این مثال در شکل ۱۰ نشان داده شده است.



شکل شماره ۱۰– مقایسه اثر روش‌های مختلف مدیریت داده‌های گم‌شده بر برآورد رابطه رگرسیونی بین متغیرها، نشان‌دهنده این‌که چگونه استفاده از رویکردهای متفاوت (حذف موارد، جای‌گذاری میانگین، جای‌گذاری رگرسیونی و جای‌گذاری چندگانه) می‌تواند منجر به برآوردها و نتایج آماری متفاوت از یک مجموعه داده واحد شود

تطبیق میانگین پیش‌بینی کننده PMM[1]

تطبیق میانگین پیش‌بینی کننده یکی از رویکردهای پرکاربرد جای‌گذاری چندگانه است که هدف آن تولید مقادیر جایگزین سازگار با توزیع واقعی داده‌ها است. در این روش، ابتدا برای هر مشاهده دارای مقدار گم‌شده، یک مقدار پیش‌بینی شده بر اساس مدل رگرسیونی مناسب محاسبه می‌شود. سپس، مجموعه‌ای از مشاهدات واقعی و غیرگم‌شده که نزدیک‌ترین مقادیر پیش‌بینی شده را دارند به‌عنوان همسایگان انتخاب می‌شوند. در نهایت، یکی از این مقادیر واقعی به‌صورت تصادفی انتخاب شده و جایگزین مقدار گم‌شده می‌شود. این رویکرد به دلیل استفاده از مقادیر واقعی موجود در داده‌ها به حفظ شکل توزیع متغیر، جلوگیری از تولید مقادیر نامعتبر و افزایش پایداری روش در برابر خطاهای مدل‌سازی کمک می‌کند. به همین دلیل PMM به طور گسترده در چارچوب MICE به‌عنوان یکی از انتخاب‌های پیش‌فرض برای داده‌های پیوسته مورد استفاده قرار می‌گیرد.

جای‌گذاری نرمال چند متغیره (MI MVN[2])

در این روش، مقادیر گم‌شده یک یا چند متغیر پیوسته با فرض وجود توزیع نرمال چند متغیره برای داده‌های کامل برآورد و جایگزین می‌شوند. در این روش، پارامترهای توزیع چندمتغیره (یعنی میانگین‌ها و ماتریس کوواریانس) با استفاده از داده‌های موجود تخمین زده شده و سپس از طریق مدل رگرسیون چندمتغیره، مقادیر گم‌شده تولید می‌گردند.

نکات عملی در جای‌گذاری چند گانه

۱. حداقل ۲۰ جای‌گذاری (m>۲۰) انجام شود تا خطای نمونه‌گیری به زیر ۵ درصد برسد.

[1] Predictive Mean Matching

[2] Impute using multivariate normal regression

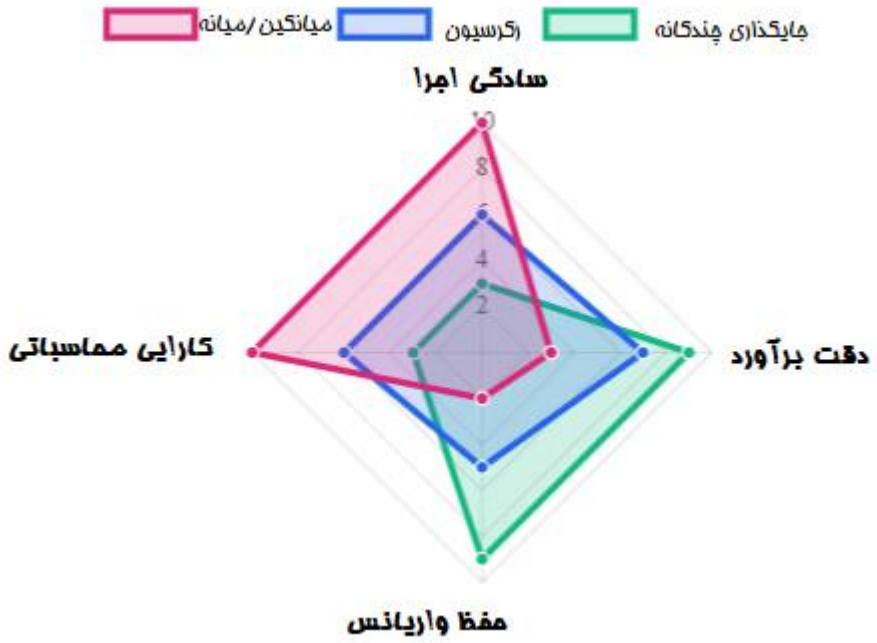
۲. تمام متغیرهای مرتبط با مکانیزم گم‌شدگی و متغیر پیامد در مدل وارد گردند تا فرض MAR قابل دفاع بماند.
۳. در روش‌های مبتنی بر الگوریتم MICE، لازم است نمودارهای زنجیره‌ای (Trace Plots) یا خلاصه آماری تکرارها بررسی شوند تا از هم‌گرایی مدل و پایداری تخمین‌ها اطمینان حاصل شود.
٤. چون MI تحت فرض MAR عمل می‌کند، بهتر است تحلیل حساسیت برای سناریوهای احتمالی MNAR انجام شود تا اثر نقض فرض MAR بر نتایج روشن شود.
۵. توزیع متغیرها، همبستگی‌ها و نمودارهای پراکنش پیش و پس از جای‌گذاری باید مقایسه شود تا از واقعی بودن مقادیر جای‌گذاری و عدم ایجاد الگوهای مصنوعی اطمینان حاصل گردد.

برای سهولت در مقایسه و مرور سریع‌تر روش‌های جای‌گذاری مطرح شده، ویژگی‌های اصلی هر روش در جدول ۲ به صورت خلاصه ارائه شده است. این جدول با سازمان‌دهی منسجم اطلاعات، امکان درک دقیق‌تر تفاوت‌های مفهومی و عملی میان روش‌ها را فراهم می‌کند.

جدول شماره ۲- خلاصه روش‌های جای‌گذاری داده‌های گم‌شده و مقایسه تعریف، مزایا و محدودیت‌های هر روش

روش جای‌گذاری	تعریف کوتاه	مزایا	محدودیت‌ها
جای‌گذاری میانگین	جایگزینی مقدار گم‌شده با میانگین همان متغیر	ساده و سریع؛ حفظ حجم نمونه	کاهش واریانس؛ تخریب همبستگی؛ احتمال سوگیری
جایگذاری میانه	جایگزینی داده گم‌شده با میانه، مناسب برای داده‌های چول	مقاوم در برابر مقادیر پرت	کاهش تنوع داده؛ تضعیف روابط آماری
جایگذاری رگرسیونی	پیش‌بینی مقادیر گم‌شده با مدل رگرسیون	استفاده از سایر متغیرها؛ دقت بیشتر	ایجاد روابط مصنوعی؛ عدم بازتاب عدم‌قطعیت
جای‌گذاری هم‌گروه (Hot Deck)	انتخاب مقدار از موارد مشابه در داده	حفظ ساختار طبیعی داده	حساس به نحوه تعریف گروه‌های مشابه
جای‌گذاری منبع ثابت (Cold Deck)	جای‌گذاری بر اساس منبع یا مقدار از پیش تعیین‌شده	مناسب برای داده‌های ثبتی	احتمال ناسازگاری با داده فعلی؛ سوگیری بالا
جای‌گذاری چندگانه(MI)	تولید چند داده جای‌گذاری شده و ترکیب نتایج	معتبرترین روش در MAR؛ انعکاس عدم‌قطعیت	پیچیده و نیازمند مدل‌سازی مناسب

همچنین به‌منظور ارائه یک مقایسه مصور میان روش‌های مختلف جای‌گذاری، نمودار رادار (شکل ۱۱) نیز ترسیم شد تا عملکرد هر روش در چهار معیار کلیدی (سادگی اجرا، دقت برآورد، حفظ واریانس و کارایی محاسباتی) به طور گویایی نمایش داده شود.



شکل شماره ۱۱- مقایسه ویژگی‌های کلیدی سه روش جای‌گذاری داده‌ها (میانگین/میانه، رگرسیون، جای‌گذاری چندگانه) بر اساس سادگی اجرا، دقت برآورد، حفظ واریانس و کارایی محاسباتی

بحث

داده‌های گم‌شده یکی از چالش های رایج و اجتناب‌ناپذیر در پژوهش‌های علوم پزشکی و اپیدمیولوژی هستند و انتخاب روش مناسب برای برخورد با آن‌ها نقش تعیین‌کننده‌ای در اعتبار نتایج دارد. مرور روش‌های متداول نشان می‌دهد که راه‌کارهای ساده مانند حذف موارد گم‌شده یا جای‌گذاری میانگین، اگرچه در ظاهر سریع و قابل اجرا هستند، در عمل می‌توانند منجر به کاهش توان آماری و سوگیری در برآوردها شوند. این محدودیت‌ها به‌ویژه زمانی تشدید می‌شود که داده‌ها تحت الگوی MAR یا MNAR گم‌شده باشند. در مقابل، روش‌های پیشرفته‌تر مانند جای‌گذاری چندگانه با الگوریتم MICE امکان لحاظ کردن عدم‌قطعیت را فراهم کرده و نسبت به رویکردهای ساده، برآوردهای پایدارتر و نزدیک‌تری به واقعیت ارائه می‌دهند. با این حال، نکته کلیدی در آموزش این است که هیچ روش واحدی برای همه موقعیت‌ها مناسب نیست و پژوهشگر باید بر اساس ویژگی‌های طراحی مطالعه، الگوی گم‌شدگی (MCAR, MAR, MNAR) و هدف تحلیل، رویکرد مناسب را انتخاب کند. در همین راستا، مقاله‌ای در NEJM[1] نیز بر این اصل تأکید دارد که هیچ روش تحلیلی واحدی قادر به رفع کامل چالش داده‌های گم‌شده نیست و همواره نیاز به تحلیل حساسیت برای بررسی استحکام یافته‌ها وجود دارد (۲).

همچنین، نتایج این مطالعه بر لزوم آموزش و آگاهی پژوهشگران از مزایا و معایب هر روش تأکید دارد. در این مقاله، روش‌های متداول برای مدیریت داده‌های گم‌شده مورد بررسی و مقایسه قرار گرفتند تا تفاوت‌ها و کارکردهای هرکدام برای پژوهشگران روشن‌تر شود. برای انتخاب مناسب‌ترین روش، پژوهشگران باید با دقت ویژگی‌های داده‌های خود را ارزیابی کرده و بر اساس آن، روش مناسب را انتخاب کنند.

محدودیت‌های مطالعه

این مقاله دارای محدودیت‌هایی است که باید به آن‌ها توجه شود:

[1] The New England Journal of Medicine

– اندازه کوچک مثال عددی که برای آموزش استفاده شد، ممکن است پیچیدگی‌های واقعی داده‌های بزرگ را منعکس نکند. مثال ارائه شده تنها شامل یک متغیر پیش‌بینی‌کننده بود و شرایط چندمتغیره پیچیده در آن نمایش داده نشد.

– تمرکز مطالعه بر روش‌های مبتنی بر MAR است و روش‌های پیچیده‌تر برای MNAR تنها به‌صورت مفهومی معرفی شده‌اند.

– عدم بررسی روش‌های مبتنی بر یادگیری ماشین نظیر (Random Forest Imputation) که امروزه کاربرد بیشتری یافته‌اند.

References

1. Graham JW. Missing data analysis: Making it work in the real world. Annual review of psychology. 2009;60(1):549–76.

2. Little RJ, D'agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. New England Journal of Medicine. 2012;367(14):1355–60.

3. van Ginkel JR, Linting M, Rippe RCA, van der Voort A. Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. J Pers Assess. 2020;102(3):297–308.

4. Abbasi R, Khajouei R, Mirzaee M. Evaluating the demographic and clinical minimum data sets of Iranian National Electronic Health Record. BMC Health Services Research. 2019;19(450):1-10.

5. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.

6. Little RJ, Rubin DB. Statistical analysis with missing data: John Wiley & Sons; 2019.

7. Allison P. Missing Data Sage Publications. Inc; 2001.

8. KAZEMI E., KARIMLO M., RAHGOZAR M.. CONTINUING EDUCATION ARTICLE: MISSING DATA. MIDDLE EASTERN JOURNAL OF DISABILITY STUDIES[Internet]. 2012;1(1):47-52.

9. Kang H. The prevention and handling of the missing data. Korean journal of anesthesiology. 2013;64(5):402–6.

10. Munsamy JI, Parrish A, Steel G. Conducting research in a resource-constrained environment: avoiding the pitfalls. InHealthcare in Low-resource Settings 2014; 2(1):14-15.

11. Graham JW, Hofer SM, MacKinnon DP. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. Multivariate behavioral research. 1996;31(2):197–218.

12. Bennett DA. How can I deal with missing data in my study? Australian and New Zealand journal of public health. 2001;25(5):464–9.

13. Little RJ. A test of missing completely at random for multivariate data with missing values. Journal of the American statistical Association. 1988;83(404):1198–202.

14. Li C. Little's test of missing completely at random. The Stata Journal. 2013;13(4):795–809.

15. Chowdhry AK, Gondi V, Pugh SL. Missing data in clinical studies. International Journal of Radiation Oncology, Biology, Physics. 2021;110(5):1267–71.

16. DeSarbo S, Green PE, Carroll JD. An alternating least-squares procedure for estimating missing preference data in product-concept testing. Decision Sciences. 1986;17(2):163–85.

17. Wisniewski SR, Leon AC, Otto MW, Trivedi MH. Prevention of missing data in clinical research studies. Biological psychiatry. 2006;59(11):997–1000.

18. Chowdhry AK, Gondi V, Pugh SL. Missing Data in Clinical Studies. Int J Radiat Oncol Biol Phys. 2021;110(5):1267–71.

19. Abbasi, R., Khajouei, R. & Mirzaee, M. Evaluating the demographic and clinical minimum data sets of Iranian National Electronic Health Record. BMC Health Serv Res 19, 450 (2019). https://doi.org/10.1186/s12913-019-4284-x.

20. Kim J-O, Curry J. The treatment of missing data in multivariate analysis. Sociological Methods & Research. 1977;6(2):215–40.

21. Barnes SA, Mallinckrodt CH, Lindborg SR, Carter MK. The impact of missing data and how it is handled on the rate of false-positive results in drug development. Pharm Stat. 2008;7(3):215–25.

22. Liu-Seifert H, Zhang S, D'Souza D, Skljarevski V. A closer look at the baseline-observation-carried-forward (BOCF). Patient Prefer Adherence. 2010;4:11–6.

23. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 2018;39(1):1–22.

24. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological methods. 2002;7(2):147-177.

Tehran University of Medical Sciences

Iranian Journal of Epidemiology

https://irje.tums.ac.ir/

IrEA Iranian Epidemiological Association

Educational Article

Missing Data in Medical Research: A Practical and Illustrated Review for Researchers and Students

Parisa Amjadi Zin Hajloo[1], Mohammad Heidari[2]

1- Master of Epidemiology, Department of Epidemiology and Biostatistics, School of Medicine, Urmia University of Medical Sciences, Urmia, Iran

2- Assistant Professor of Epidemiology, Department of Epidemiology and Biostatistics, School of Medicine, Urmia University of Medical Sciences, Urmia, Iran

DOI:

Article Information

Received
06 September 2025

Accepted
05 January 2026

Corresponding author
Mohammad Heidari

Corresponding author E-mail
heidari.m@umsu.ac.ir

Keywords:
Missing data, Missing data patterns, MCAR, MAR, MNAR, Multiple imputation

Abstract

Missing data is a common and unavoidable challenge in medical and epidemiological research, often leading to biased estimates, reduced statistical power, and misleading interpretations when not properly addressed. Despite its importance, accessible and practical educational resources on this topic remain limited in Persian. This educational article provides a clear and structured overview of the fundamental concepts of missing data, including definitions, common patterns (univariate and multivariate), and the three major mechanisms of missingness: MCAR, MAR, and MNAR. A range of widely used approaches for handling missing data is summarized, from basic methods such as case deletion and simple imputation to more advanced techniques including multiple imputation and likelihood-based procedures (EM and MLE). Practical examples and visual illustrations are incorporated to facilitate conceptual understanding. The ultimate goal of this article is to provide a practical framework for researchers and students, enabling them to select the appropriate approach for dealing with missing data in the design and analysis of their research and to prevent analytical errors.