انواع گمشدگی در مطالعات طولی و روش های مبنی بر درستنمایی برای تحلیل آن ها

فرید زایری [*1]، علی رضا اکبرزاده باغبان [2]، مژگان کاظم زاده [3]، مهدی یاسری [4]، علی محمد عباسی [5]

1) مرکز تحقیقات پروتئومیکس، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی
2) گروه علوم پایه، دانشکده توانبخشی، دانشگاه علوم پزشکی شهید بهشتی
3) گروه آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی
4) گروه اپیدمیولوژی و آمارزیستی،دانشکده بهداشت، دانشگاه علوم پزشکی تهران
5) گروه بهداشت حرفه ای، دانشکده بهداشت، دانشگاه علوم پزشکی ایلام

تاریخ دریافت: 91/8/17 **تاریخ پذیرش: 91/11/23**

چکیده

وجود مقادیر گمشده در مجموعه داده ها به وفور در پژوهش های علوم مختلف از جمله علوم پزشکی و به ویژه در مطالعات طولی، که در آن ها هر فرد در طول زمان تحت اندازه گیری های مکرر قرار می‌گیرد، به چشم می خورد. در چند دهه اخیر فعالیت های آماری وسیعی در این زمینه، در حوزه های مفاهیم، مباحث، روش های تئوری و نرم افزاری، انجام شده است. با وجود گسترش استفاده از نتایج حاصل از این فعالیت ها، در بسیاری از موارد شاهد برداشت های مبهم محققین از مفاهیم و در نتیجه استنباط های ناصحیح هستیم. بنا بر این با توجه به اهمیت مساله و نیاز جامعه علمی به آشنایی صحیح و دقیق با مباحث داده های گمشده، در مقاله حاضر به مرور و مقایسه مفاهیمی مانند الگوها و مکانیزم های گمشدگی و مدل های موجود برای تحلیل داده های طولی با مقادیر گمشده پرداخته شده است. علاوه بر این کاربرد این مدل ها در داده های یک کارآزمایی بالینی ترک سیگار با متغیر پاسخ پیوسته نشان داده خواهند شد.

واژه‌های کلیدی: گمشدگی کاملاً تصادفی، گمشدگی تصادفی، گمشدگی غیرتصادفی، مدل گزینش، مدل الگوی آمیخته، مدل پارامترهای مشترک

* **نویسنده مسئول:** مرکز تحقیقات پروتئومیکس، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی

Email:

مقدمه

در مطالعات طولی، هر فرد یا واحد آزمایش تحت اندازه گیری اولیه و اندازه گیری های مکرر در طول زمان قرار می‌گیرد. در چنین مطالعاتی وجود داده های ناکامل یا اصطلاحاً داده های گمشده(missing data) امری اجتناب ناپذیر است؛ زیرا ممکن است تعدادی از افراد به دلایل مختلف در تمام زمان های اندازه گیری در دسترس نباشند. چنین داده هایی چالش هایی در تحلیل ها و مدل سازی های آماری به وجود می‌آورد.

به طور کلی می توان چهار گروه از استراتژی ها را برای برخورد با داده های گمشده در مطالعات طولی در نظر گرفت. اولین و ساده‌ترین استراتژی، شامل حذف واحدهای دارای داده های ناقص از مطالعه و انجام تحلیل های آماری با تکیه بر اطلاعات واحدهایی است که در تمام زمان های اندازه گیری در دسترس بوده‌اند (complete-case analysis). متاسفانه در حالی که استفاده از این روش به دلیل سادگی آن در بین محققین بسیار رایج است؛ در اکثر مواقع منجر به نتایج اریب و نامعتبر می‌شوند. استراتژی های دیگر شامل روش های جانهی(imputation)،(3-1)، روش های وزن دهی،(۴،۵)، و روش های مبنی بر تابع درستنمایی هستند. در حالت کلی هر سه روش مذکور به جانهی مقادیر گمشده با مقادیری خاص می پردازند؛ با این تفاوت که جانهی مقادیر گمشده در روش اول به طور واضح و مستقیم ولی در دو روش دیگر به طور ضمنی و غیرمستقیم انجام می‌شود(منظور از جانهی، جایگذاری مقادیری معقول به جای مقادیر گمشده است).(6)

مشروط به برقراری فرضیات متفاوت در مورد الگو و مکانیزم داده های گمشده، روش مورد استفاده نیز متفاوت خواهد بود. فعالیت های آماری انجام گرفته در معرفی و بررسی مزایا و معایب هر یک از استراتژی های فوق بسیار گسترده است به طوری که پرداختن به همه این مباحث در یک مقاله غیرممکن است. تمرکز مقاله حاضر بر مرور و مقایسه روش های مبتنی بر درستنمایی است.

روبین در سال 1976 سه مکانیزم مهم برای گمشدگی داده ها معرفی کرد،(1). وقتی گمشدن هیچ ارتباطی با داده ها نداشته باشد، گمشدگی کاملاً تصادفی(missing completely at random یا MCAR) رخ می‌دهد. چنان چه گمشدن تنها بستگی به داده های مشاهده شده داشته باشد ولی به داده های مشاهده نشده وابسته نباشد، گمشدگی تصادفی (missing at random یا MAR) نامیده می‌شود و در حالتی که مکانیزم گمشدن به داده های مشاهده نشده نیز وابسته باشد، با گمشدگی غیرتصادفی (missing not at random یا MNAR) مواجه هستیم.(7)

تحت برقراری شرایطی، مکانیزم های MCAR و MAR گمشدگی های قابل چشم پوشی(ignorable) نامیده می‌شوند. در این صورت گمشدگی ها، قابل اغماض و چشم پوشی و استفاده از روش های معمول تحلیل داده های طولی نامتعادل مانند مدل های حاشیه ای،(8)، مدل های اثرات تصادفی،(9)، و مدل های انتقال،(10)، برای تحلیل داده ها مناسب خواهد بود.(11)

در حالی که استفاده از روش های فوق در مقابل روش های ساده ای چون تحلیل مبنی بر داده های کامل، گسترش پیدا کرده است، در بسیاری از مطالعات فرض قابل چشم پوشی بودن گمشده ها برقرار نیست؛ بلکه در مقابل با گمشدگی های غیرقابل چشم پوشی (MNAR) مواجه هستیم. در چنین مواردی، استفاده از مدل های استاندارد فوق، منجر به برآوردهای اریب و نامعتبر خواهد شد. بنا بر این نیازمند مدل سازی های پیشرفته آماری برای تحلیل چنین داده هایی هستیم. در سال های اخیر سه گروه از مدل ها در این زمینه معرفی شده و گسترش پیدا کرده‌اند: مدل های گزینش (selection models)، مدل های الگوی آمیخته (pattern-mixture models) و مدل های پارامترهای مشترک(shared-parameters models).(12)

در مقاله حاضر ابتدا به طور مفصل مفاهیم اشاره شده در این قسمت شامل گمشدگی در داده های طولی و انواع آن بیان می‌شوند. در بخش های بعد، به معرفی مدل های گزینش، الگوی آمیخته و پارامترهای مشترک برای کنترل گمشدگی های غیرقابل چشم پوشی و بررسی مزایا و معایب هر یک خواهیم پرداخت. علاوه بر این برای نشان دادن کاربرد این

مدل ها به تحلیل داده های کارآزمایی بالینی ترک سیگار،(13)، با متغیر پاسخ پیوسته خواهیم پرداخت.

گمشدگی و مشکلات ناشی از آن در مطالعات طولی

استفاده از مطالعات طولی و کارآزمایی های بالینی در علوم مختلف از جمله علوم پزشکی بسیار گسترده است. مشکل عمده استفاده از این طرح ها، وجود داده های گمشده است. هرگاه یک یا چند دنباله از اندازه های مکرر مربوط به واحدها یا افراد درون مطالعه ناکامل باشد، با مقادیر گمشده در مجموعه داده ها مواجه هستیم،(14)،(هر چند مفهوم گمشدن ساده به نظر می رسد ولی در مواردی در نظر گرفتن داده های ثبت نشده به عنوان گمشده بی معنا خواهد بود. توجه به این نکته ضروری است که منظور از داده گمشده مقداری است که بر اساس طرح مطالعه باید ثبت می شده ولی به دلیلی ثبت نشده است. بنا بر این چنان چه طراحی مطالعه به گونه ای باشد که برای بعضی از افراد زمان های بیشتری برای ثبت اندازه های مکرر و برای بعضی دیگر زمان های کمتر در نظر گرفته باشد، با وجود این که مجموعه داده نامتعادل خواهد بود، گمشدگی رخ نداده است،(14). علاوه بر این در مطالعاتی مثل بررسی کیفیت زندگی، در نظر گرفتن داده های ثبت نشده افراد فوت شده در حین مطالعه، به عنوان داده های گمشده بی معنا است،(15). حضور این مقادیر دلایل مختلفی می‌تواند داشته باشد از جمله در دسترس نبودن افراد، کناره گیری افراد از مطالعه، بروز اثرات جانبی منفی، موثر نبودن مداخله، شرایط سخت درمان و ... ،(16). به طور کلی حضور مقادیر گمشده در داده ها سه مشکل عمده ایجاد می‌کنند: نخست این که قسمتی از اطلاعات از دست می رود، در نتیجه با کاهش حجم نمونه و متعاقباً با کاهش دقت و کارایی برآوردها رو به رو هستیم. از طرف دیگر حضور این مقادیر منجر به ایجاد مجموعه داده ای نامتعادل می شود، زیرا همه افراد دارای تعداد یکسانی از اندازه گیری های مکرر در زمان های یکسان نخواهند بود. بنا بر این امکان استفاده مستقیم از مدل ها و روش هایی که برای داده های کامل قابل اجرا هستند، وجود ندارد. علاوه بر موارد فوق، از آن جا که ممکن است اطلاعات گمشده مربوط به افرادی باشد که تفاوت های اساسی با افراد دیگر(افراد با اطلاعات کامل) داشته باشند، پتانسیل اریبی وجود دارد. برطرف کردن اریبی در این موارد مشکل است زیرا دلایل دقیق گمشدگی معمولاً ناشناخته‌اند.(۱۶)

گمشدگی در مطالعات طولی به طور کلی می تواند در پیامد(متغیر پاسخ) مورد نظر و یا هر یک از متغیرهای مستقل وابسته به زمان رخ دهد. مباحث مطرح شده در این مقاله بر گمشدگی در متغیر پاسخ تمرکز دارند. این مباحث گاهی مستقیماً و گاهی با اصلاحاتی اندک قابل تعمیم به گمشدگی در متغیرهای مستقل نیز هستند.

مکانیزم داده های گمشده (missing data mechanisms)

بحث گمشدگی در مطالعات طولی حساس تر از مطالعات مقطعی است زیرا گمشدگی می تواند در اشکال و موقعیت های مختلفی رخ دهد،(1). به همین دلیل در تحلیل داده های طولی با گمشدگی، معمولاً نیاز به فرضیاتی درباره دلایل گمشدگی داریم که مکانیزم داده های گمشده نامیده می‌شوند. در حالت کلی وقتی می توانیم استنباط های معتبر در مورد داده های ناکامل داشته باشیم که مکانیزم داده های گمشده به طور کامل شناخته شده باشند ولی از آن جایی که این مکانیزم تحت کنترل محقق نیست، شناخت دقیق آن امکان پذیر نیست. بنا بر این در هر مطالعه فرضیاتی درباره این مکانیزم پذیرفته می شود. این فرضیات در واقع در پاسخ به این سوال که «چرا گمشدگی رخ داده و یا به طور خاص تر، آیا مقادیر گمشده ارتباطی با سوال های کاربردی تحقیق دارند یا نه؟» شکل می گیرند. چنان چه دلایل گمشدگی با پیامد مورد نظر مرتبط نباشند، مشکل خاصی در تحلیل داده ها وجود نخواهد داشت. ولی در صورتی که این ارتباط وجود داشته باشد، ممکن است منجر به برآوردهای اریب شود؛ زیرا داده های در دسترس با داده های کامل متفاوت خواهند بود. مسلماً اعتبار تحلیل ها به برقراری یا عدم برقراری این فرضیات وابسته است.(۶،۱۴)

روبین در سال 1976 و لیتل و روبین در سال 1987، راهکارهایی برای تحلیل آماری داده های طولی

ناقص ارائه دادند که شامل دسته بندی مفیدی از مکانیزم های گمشدگی بود،(17). فرض کنید N تعداد کل افراد نمونه و J تعداد مشاهدات تکراری برای هر فرد را نشان دهد. بردار پاسخ برای فرد i ام به صورت $y_i=(y_{i1},\ldots, y_{ij})$ خواهد بود. برای هر بردار پاسخ برداری از متغیرهای نشانگر به صورت $r_i=(r_{i1},\ldots, r_{ij})$ تشکیل می‌دهیم که در آن $r_{ij}=0$ در صورتی که y_{ij} مشاهده شده باشد و $r_{ij}=1$ در صورتی که y_{ij} مشاهده نشده باشد. با تعریف بردار r_i، بردار مشاهدات برای هر فرد به صورت y_i^{obs} و y_i^{mis} تفکیک می‌شود. که مولفه اول مربوط به مقادیر گمشده و مولفه دوم مربوط به مقادیر مشاهده شده است.

مکانیزم گمشدگی در حقیقت احتمال مشاهده یا عدم مشاهده پاسخ در هر زمان را مشخص می‌کند و به صورت یک مدل احتمالی برای توزیع نشانگرهای پاسخ r_i به شرط y_i^{obs} و y_i^{mis} بیان می‌شود. با در نظر گرفتن نوع ارتباط r_i با y_i، سه نوع مکانیزم گمشدگی خواهیم داشت:

1-گمشدگی کاملاً تصادفی(MCAR)

این نوع گمشدگی زمانی اتفاق می‌افتد که احتمال مشاهده نشدن پاسخ در یک زمان به هیچ یک از پاسخ های مشاهده شده و مشاهده نشده(قبلی و بعدی) بستگی نداشته باشد، یعنی:

$$P(r_i|y_i^{obs}, y_i^{mis}, X_i) = p(r_i|X_i)$$

برای روشن شدن موضوع موقعیتی را در نظر بگیرید که افراد شرکت کننده در مطالعه در مناطق مختلف شهری ساکن هستند. چنان چه دلیل عدم ثبت داده های افراد، مسائلی چون بارش برف یا ترافیک سنگین در روز تعیین شده برای مراجعه به مرکز جهت اندازه گیری باشد، گمشدگی کاملاً تصادفی رخ داده است. زیرا گم شدن داده ها ارتباطی با مشاهدات قبلی بردارy_i ندارد. در حقیقت در مکانیزم MCAR، داده های مشاهده شده نمونه ای تصادفی از کل داده ها خواهند بود.

2-گمشدگی تصادفی(MAR)

گمشدگی تصادفی زمانی اتفاق می‌افتد که احتمال گم شدن پاسخ در یک زمان، به پاسخ های مشاهده شده وابسته باشد ولی به پاسخ های مشاهده نشده بستگی نداشته باشد. یعنی:

$$P(r_i|y_i^{obs}, y_i^{mis}, X_i) = p(r_i|y_i^{obs}, X_i)$$

چنان چه مشاهده می‌شود در این نوع از گمشدگی، احتمال گم شدن ممکن است به مشاهدات قبلی وابسته باشد. این مکانیزم نسبت به مکانیزم MCAR معقول تر به نظر می رسد، هر چند باعث ایجاد محدودیت های بیشتر در تحلیل خواهد شد، زیرا در مکانیزم MAR پاسخ های مشاهده شده، نمونه ای تصادفی از کل داده ها نخواهد بود.

گمشده های ساختاری(missing by design) نمونه‌ای از گمشده های MAR هستند. برای مثال ممکن است منشور(protocol) یک مطالعه طوری طراحی شود که در صورتی که مقدار متغیر پاسخ برای فردی خارج از محدوده کلینیکی خاصی ثبت شود، آن فرد از مطالعه خارج شود،(15). در چنین مواردی، گمشدگی در y_i تحت کنترل محقق است و تنها به مولفه‌های مشاهده شده y_i بستگی دارد.

3-گمشدگی غیر تصادفی(MNAR)

گمشدگی غیرتصادفی وقتی رخ می‌دهد که احتمال مشاهده نشدن پاسخ در یک زمان، حداقل به یکی از پاسخ هایی که باید مشاهده می شده اند ولی گم شده اند وابسته باشد. یعنی $P(r_i|y_i^{obs}, y_i^{mis}, X_i)$ حداقل به یکی از مولفه‌های y_i^{mis} وابسته است.

برای مثال مطالعه ای را در نظر بگیرید که هدف آن اندازه گیری کیفیت زندگی افراد است(متغیر پاسخ). چنان چه افرادی که سطح کیفیت زندگی بالا یا پایین برخوردارند، از پر کردن پرسش نامه و کامل کردن مطالعه اجتناب کنند، گمشدگی غیرتصادفی رخ داده است. به عنوان مثالی دیگر مطالعه ای را در نظر بگیرید که بر روی بیماران تنفسی در شهری خاص صورت می گیرد و علت عمده گمشدگی داده های آن، مهاجرت بیماران باشد. چنان چه علت مهاجرت، گسترش مشکل تنفسی افراد به دلیل آلودگی هوای آن شهر باشد، گمشدگی از نوع MNAR خواهد بود؛ زیرا مکانیزم گمشدگی به مقدار گمشده وابسته است.

2-2-گمشدگی های قابل چشم پوشی و غیرقابل چشم پوشی

مفهوم قابل چشم پوشی بودن برای اولین بار در سال 1976 توسط روبین همراه با معرفی مدل های گزینش مطرح شد و پس از آن به طور گسترده مورد استفاده محققین قرار گرفت. تحت برقراری دو شرط زیر، گمشده ها قابل چشم پوشی خواهند بود: الف) مکانیزم گمشدگی MCAR و یا MAR باشد. به عبارت دیگر r_i حداقل مستقل از y_i^{mis} باشد(مشروط بر y_i^{obs} و X_i). ب) پارامترهای فرایند اندازه‌گیری(θ) مستقل از پارامترهای مکانیزم گمشدگی(φ) باشند. در روش های مبنی بر درستنمایی، با برقرار بودن فرض قابل چشم پوشی بودن تابع لگ درستنمایی برای θ و همین تابع برای φ می توانند از هم مجزا شوند؛ یعنی:

$$l(\theta, \varphi| y_i^{obs}, r_i) = l(\theta| y_i^{obs}) + l(\varphi | y_i^{obs}, r_i)$$

بیشینه سازی این رابطه مستلزم حداکثر سازی جداگانه‌ی دو عبارت سمت راست است،(6)، چون عبارت

$$l(\varphi | y_i^{obs}, r_i)$$

شامل هیچ اطلاعی درباره θ نیست، می توانیم در مواردی که می خواهیم درباره θ استنباط کنیم، آن را نادیده بگیریم. علت نام گذاری«قابل چشم پوشی بودن» برای دو مکانیزم MCAR و MAR در این شرایط نیز همین مساله است. در مقابل، گمشدگی های غیرتصادفی(MNAR)، گمشدگی های غیرقابل چشم پوشی یا آگاهی بخش(informative) نامیده می شوند. لازم به توضیح است که چنان چه فضای پارامترهای θ و φ مشترک باشند، نادیده گرفتن عبارت $l(\varphi | y_i^{obs}, r_i)$ منجر به از دست رفتن کارایی می‌شود. علاوه بر این در چارچوب استنباط های فراوانی گرا، تنها مکانیزم MCAR قابل چشم پوشی خواهد بود.(6)

3-2-الگوهای داده های گمشده (patterns of missing data)

علاوه بر مکانیزم های گمشدگی، تشخیص الگوی گمشدگی نیز، در انتخاب روش مناسب برای تحلیل داده های طولی حائز اهمیت است. دو الگوی رایج گمشدگی در مطالعات طولی الگوی یکنواخت یا انصراف(monotone یا dropout) و الگوی غیریکنواخت یا متناوب(non-monotone یا informative) هستند. الگوی انصراف به مواردی اطلاق می‌شود که در یک زمان خاص از مطالعه خارج شده اند و دیگر بازنگشته اند. به عبارتی چنان چه مولفه y_{ik} از بردار y_i گمشده باشد، تمام مشاهدات بعدی یعنی $y_{in},\ldots,y_{ik+1}$ نیز گمشده اند که در نتیجه با الگویی یکنواخت در داده های گمشده مواجه هستیم. در مقابل اگر داده های مربوط به حداقل یکی از شرکت کنندگان به طور غیریکنواخت ثبت شده باشد، الگوی گمشدگی به صورت متناوب رخ داده است. برای مثال فردی را در نظر بگیرید که در یک یا چند دوره پیگیری برای اندازه‌گیری مراجعه نکرده ولی در دوره های بعدی در مطالعه حضور داشته است.

با وجود این‌که الگوهای متناوب می‌توانند به اشکال مختلف در مجموعه داده ها مشاهده و منجر به تشکیل توابع درستنمایی پیچیده شوند، از نظر مفهومی معمولاً کنترل این نوع از گمشده ها ساده تر از گمشده های ریزش شده است، زیرا افراد دارای الگوی متناوب به طور کامل از دسترس خارج نشده اند و حضور آنان در ادامه مطالعه به طور کلی دلیل مناسبی برای چشم پوشی از اطلاعات از دست رفته آنان خواهد بود. در مقایسه، ارزیابی توابع درستنمایی در الگوهای گمشدگی یکنواخت ساده تر است زیرا این توابع به صورت شرطی قابل تجزیه اند. ولی بر خلاف گمشدگی های متناوب، در این موارد چشم پوشی از انصراف به سادگی میسر نخواهد بود؛ زیرا معمولاً با کناره گیری افراد از مطالعه رو به رو هستیم و ممکن است این مساله به طور مستقیم یا غیرمستقیم با پیامد مورد اندازه‌گیری در ارتباط باشد. بنا بر این الگوی انصراف از اهمیت بیشتری نسبت به الگوی متناوب برخوردار است. موضوع کلیدی در بحث انصراف، بررسی تفاوت افراد خارج شده از مطالعه و باقی مانده در مطالعه است. چنان چه این دو گروه متفاوت نباشند، تحلیل های محدود شده به اطلاعات افراد باقی مانده در مطالعه معتبر خواهد بود، هر چند ممکن است کارایی کاهش یابد. در غیر این صورت، چنین تحلیل هایی به طور بالقوه اریب خواهند بود.(۱،۱۶)

بر اساس طبقه بندی لیتل و روبین برای مکانیزم داده های گمشده، می‌توان سه نوع انصراف نیز در نظر گرفت. اگر احتمال انصراف در هر زمان اندازه گیری، مستقل از همه پاسخ های قبلی، پاسخ فعلی و پاسخ های آینده باشد(به شرط متغیرهای مستقل)، انصراف به صورت کاملاً تصادفی رخ داده است. در صورتی که احتمال انصراف در هر مراجعه، به شرط پاسخ های مشاهده شده، مستقل از پاسخ فعلی و پاسخ های آینده باشد، انصراف از نوع تصادفی خواهد بود. در نهایت چنان چه احتمال ریزش در هر زمان، وابسته به پاسخ مشاهده نشده فعلی یا پاسخ های مشاهده نشده آتی باشد، انصراف غیرتصادفی خواهیم داشت. متعاقباً انصراف می تواند آگاهی بخش یا غیرآگاهی بخش(قابل چشم پوشی) باشد. برای مثال دو شخص با پاسخ های یکسان تا زمان t را در نظر بگیرید که یکی انصراف داده و دیگری در مطالعه باقی مانده باشد. در شرایط MAR، توزیع مشاهدات آتی برای هر دو نفر یکسان خواهد بود؛ در حالی که چنان چه انصراف از نوع NAR باشد، حاکی از توزیع های متفاوت مشاهدات آتی برای آن دو نفر است.(۶،۷،۱۸)

3 استنباط مبنی بر درستنمایی با حضور داده های گمشده

با توجه به مباحث فوق و اهمیت مکانیزم آگاهی بخش و الگوی انصراف، در ادامه پس از بحث مختصری بر روش های تحلیل گمشده های قابل چشم پوشی، به معرفی و مقایسه مدل های موجود برای تحلیل انصراف آگاهی بخش خواهیم پرداخت.

3-1-مدل های طولی برای گمشدگی های قابل چشم پوشی

به طور معمول 3 دسته از مدل ها برای تحلیل داده های طولی کاربرد دارند که هر سه گسترشی از مدل های خطی تعمیم یافته،(19)، هستند:

1-مدل های حاشیه ای

2-مدل های آمیخته با اثرات تصادفی

3-مدل های انتقال

مدل های خطی تعمیم یافته و سه مدل فوق در منابع زیادی مورد بحث و نقد قرار گرفته اند. پرداختن به این موارد از حوصله مقاله حاضر خارج است؛ خواننده علاقمند می‌تواند در این زمینه به منابع دیگر رجوع کند.(20-23)

یکی از مزیت های عمده این مدل ها نسبت به روش های ساده ای چون تحلیل واریانس اندازه های تکراری، امکان برازش آن ها با تعداد داده های نابرابر برای افراد مختلف در مطالعات طولی است. روش برآورد پارامترها در هر سه مدل فوق می‌تواند بر مبنای روش حداکثر درستنمایی باشد ولی در مدل های حاشیه ای معمولاً از روش معادلات برآوردی تعمیم یافته(Generalized Estimation Equations یا GEE)،(24)، که بر پایه روش شبه درستنمایی استوار است، استفاده می‌شود.

وقتی گمشده ها مکانیزم MCAR دارند، افراد با داده های ناقص در حقیقت زیرمجموعه ای تصادفی از نمونه هستند و بنا بر این مقادیر مشاهده شده پاسخ نیز یک زیرنمونه تصادفی از تمام مقادیر پاسخ خواهند بود. در این صورت هر روشی برای تحلیل در نظر گرفته شود(فقط بر اساس بردارهای کامل (complete-cases) یا بر اساس داده های در دسترس(available-cases)، برآوردها اریب نخواهند بود. در واقع تمام روش های فوق، به طور معتبر روند میانگین پاسخ ها و ارتباطات درون شخصی را برآورد می‌کنند.(18)

در مقابل وقتی مکانیزم گمشده ها MAR باشد، افراد با داده های ناقص زیرمجموعه تصادفی از نمونه نخواهند بود؛ تنها چنان چه افراد بر حسب مقادیر مشاهده شده پاسخشان طبقه‌بندی شوند(یعنی بر حسب y_i^{obs}) می توان آن ها را به عنوان زیرمجموعه ای تصادفی از نمونه های آن طبقه خاص در نظر گرفت. بنا بر این مقادیر مشاهده شده نیز لزوماً زیرنمونه تصادفی از پاسخ ها نیستند. به طور کلی، توزیع y_i^{obs}، یعنی مولفه های مشاهده شده y_i، از توزیع مولفه های مشابه y_i در جمعیت هدف(جامعه)، متفاوت است. این نشان می دهد که میانگین و کواریانس نمونه ای که صرفاً بر اساس داده های کاملیا داده های در دسترس به دست می‌آیند، برآوردهای اریبی برای میانگین و کواریانس جامعه خواهند بود.(18)

وقتی گمشده ها MAR هستند(و البته MCAR نیستند) روش های بر مبنای بردارهای کامل و یا روش GEE استاندارد، برآوردهای اریب از میانگین می دهند. در مقابل روش های بر مبنای درستنمایی که به طور دقیق توزیع توام کامل پاسخ ها را مشخص می‌کنند، برآوردهای معتبر می‌دهند. البته یک شرط ظریف ولی مهم وجود دارد و آن این است که توزیع توام کامل باید به طور صحیح مشخص شود. در عمل این به این معناست که نه تنها مدل میانگین پاسخ باید به طور صحیح مشخص شود، بلکه مدل ارتباط درون فردی نیز باید درست انتخاب شود. بنا بر این وقتی گمشدگی MAR است، روش های بر مبنای درستنمایی، به شرط این که ماتریس کواریانس به درستی مدل شود، استنباط های معتبر به دست می دهند. به طور مشابه مدل های اثرات تصادفی نیز به شرط این که ساختار اثرات تصادفی به طور صحیح مشخص شوند، برآوردهای معتبر برای اثرات ثابت نتیجه خواهند داد. به طور خلاصه، وقتی گمشدگی MAR باشد، استنباط راجع به میانگین به هر نوع اشتباه در تعیین توزیع توام بردار پاسخ، بسیار حساس است. بنا بر این وقتی داده های طولی ناقص هستند در تعیین ارتباط درون فردی باید دقت کرد.(۱۱،۱۸)

در روش GEE استاندارد لازم است که مدلی برای میانگین مشاهدات به شرط متغیرهای پیشگو داشته باشیم. در MAR، این مدل به طور کلی برای داده های مشاهده شده برقرار نیست، بنا بر این اعتبار تحلیل ها به خطر می افتد. با استفاده از یک GEE وزنی ساده،(۲۲،۲۵)، روش هایی برای تعدیل تحلیل ها ارایه می‌شود که در آن وزن ها با استفاده از یک مدل برای احتمال گمشدگی برآورد می شوند. بنا بر این مدل گمشدگی باید به طور دقیق مشخص و برآورد شود. برای اطلاعات بیشتر در این زمینه می‌توان به منبع دیگری مراجعه نمود.(18)

در نهایت وقتی گمشدگی NMAR است، تقریباً هیچ یک از روش‌های استاندارد تحلیل داده های طولی معتبر نیستند. هم روش های GEE و هم روش های بر مبنای درستنمایی به برآوردهای اریب از میانگین پاسخ منجر می‌شوند،(18). بنا بر این روش های مدل سازی پیشرفته برای گمشده های غیرقابل چشم پوشی مورد نیاز است.

قبل از ورود به بحث گمشده های آگاهی بخش، به ارائه مثالی شبیه سازی شده در راستای مفاهیم فوق می پردازیم:

3-2- *شبیه سازی MCAR، MAR و MNAR*

هدکر و گیبنز در سال 2008 به مطالعه ای شبیه سازی شده جهت تشخیص روش های مناسب کنترل داده های گمشده در مطالعات طولی پرداختند،(26). داده ها بر اساس مدل زیر شبیه سازی شدند:

$$y_{ij}=\beta_0+\beta_1 time_j+\beta_2 group_i+\beta_3(group_i\times time_j)+v_0+v_{1i}time_j+\varepsilon_{ij}$$

به طوری که $time_j$ شامل پنج نقطه زمانی 0، 1، 2، 3، 4 و $group_i$ یک متغیر دو حالتی با مقادیر 0 یا 1 بودند. ضرایب رگرسیونی به صورت $\beta_0=25$، $\beta_1=-1$، $\beta_2=0$ و $\beta_3=-1$ در نظر گرفته شدند. اثرات تصادفی افراد(v_0 و v_1) دارای توزیع نرمال با میانگین صفر و واریانس های $\sigma^2_{v0}=4$، $\sigma^2_{v1}=25$ و کواریانس $\sigma_{v01}=-1$ و خطاهای ε_i دارای توزیع نرمال با میانگین صفر و واریانس $\sigma^2=4$ فرض شدند. ابتدا 5000 داده با پنج زمان اندازه گیری بر اساس مدل و پارامترهای ذکر شده تولید شد. سپس به تولید داده های گمشده با مکانیزم های MCAR، MAR و MNAR پرداختند. در نهایت برای تحلیل داده ها مدل اثرات تصادفی با عرض از مبدا تصادفی و روند زمانی و نیز مدل حاشیه ای با برآوردهای GEE استاندارد برازش داده شد. نتایج حاصل در جدول شماره 1 آمده است.

جدول شماره 1. نتایج شبیه سازی MCAR، MAR و MNAR؛ برآوردها(انحراف معیارها)

	مدل	β_0	B_1	β_2	B_3	σ^2_{v0}	σ_{v01}	σ^2_{v1}	σ^2
مقادیر شبیه سازی	---	25	-1	0	-1	4	-۰،۱	۰،۲۵	4
داده های کامل	اثرات تصادفی	۲۴،۹۶۹	-۰،۹۹۴	-۰،۰۰۱	-۰،۹۸۶	۳،۹۱۸	-۰،۰۲۰	۰،۱۹۹	۴،۰۷۰
		(۰،۰۵۰)	(۰،۰16)	(۰،۰۷۱)	(۰،۰۲۳)	(۰،۱۲۹)	(۰،۰۳۲)	(۰،۰۱۴)	(۰،۰۴۶)
MCAR	اثرات تصادفی	۲۴،۹۹۱	-۱،۰۲۴	-۰،۰۸۷	-۰،۹۳۳	۳،۸۱۱	-۰،۰۲۰	۰،۱۹۹	۴،۰۷۰
		(۰،۰۶۳)	(۰،۰۲۳)	(۰،۰۸۹)	(۰،۰۳۲)	(۰،۱۹۳)	(۰،۰۵۶)	(۰،۰۲۵)	(۰،۰۸۳)
MAR	اثرات تصادفی	۲۴،۹۹۶	-۱،۰۳۹	-۰،۰۱۰	-۰،۹۶۹	۳،۹۸۱	-۰،۰۶۴	۰،۲۳۳	3،۸۷۳
		(۰،۰۵۳)	(۰،۰۲۵)	(۰،۰۷۵)	(۰،۰۴۱)	(۰،۱۵۸)	(۰،۰۶۵)	(۰،۰۳۲)	(۰،۰۷۸)
	حاشیه‌ای	25,281	-۱،۱۶۴	۰،۰۱۹	-۱،۰۰۱				
		(۰،۰۵۸)	(۰،۰۳۷)	(۰،۰۹۷)	(۰،۰۸۵)				
MNAR	اثرات تصادفی	۲۴،۹۵۶	-۰،۲۳۳	۰،۰۲۷	-۰،۵۵۲	۳،۸۵۶	-۰،۹۴۳	۰،۳۱۹	۳،۰۲۰
		(۰،۰۴۹)	(۰،۰۲۰)	(۰،۰۷۰)	(۰،۰35)	(۰،۱۳۱)	(۰،۰۵۱)	(۰،۰۲۵)	(۰،۰۵۳)
	حاشیه ای	۲۵،۰۵۱	-۰،۳۸۶	۰،۰۱۶	-۰،۵۸۳				
		(۰،۰۴۹)	(۰،۰۲۰)	(۰،۰۷۱)	(۰،۰۳۴)				

با توجه به جدول شماره 1 مشاهده می‌شود که در حالت MCAR، برآوردهای به دست آمده از برازش مدل اثرات تصادفی به خوبی مقادیر واقعی پارامترها را پوشش می‌دهند، البته خطاهای استاندارد در حضور داده های گمشده بزرگ تر از همین خطاها بدون حضور داده های گمشده است که این مساله اثر بروز گمشدگی در مجموعه داده را نشان می‌دهد. از برازش مدل حاشیه ای با GEE استاندارد نیز نتایج مشابه به دست آمد که در جدول گزارش نشده است.

اطلاعات جدول شماره 1 هم چنین نشان می‌دهد که در حالت MAR مدل اثرات تصادفی برآوردهایی معقول از پارامترها به دست می‌دهد، در حالی که برازش مدل حاشیه ای منجر به چنین برآوردهایی نمی‌شود. تولومی و همکاران نیز در مطالعه‌ای مشابه نشان دادند که اریبی ناشی از برازش مدل حاشیه ای با GEE استاندارد، با افزایش درصد و شدت غیرتصادفی بودن گمشدگی، افزایش می‌یابد.

نتایج مربوط به حالت MNAR نیز در جدول شماره 1، نشان می دهد که هیچ یک از مدل های اثرات تصادفی و حاشیه ای در برآورد پارامترها موفق نبوده‌اند. بنا بر این در شرایط MNAR، استفاده از مدل های تحلیل داده های MCAR و MAR مشکل ساز خواهد بود.

در دهه های اخیر، گسترش روش هایی برای رویارویی با چنین داده های گمشده ای، به شدت مورد توجه قرار گرفته است. فعالیت های انجام شده در این زمینه بیشتر بر اساس مدل سازی توام نشانگرهای گمشدگی و مقادیر اندازه های تکراری(شامل مشاهده شده ها و گمشده ها) بوده است. در ادامه به معرفی سه دسته از این مدل ها خواهیم پرداخت.

3-2-مدل های توام برای گمشدگی های غیرقابل چشم پوشی

وقتی بر این باوریم که قابل چشم پوشی بودن فرض مناسبی برای گمشده ها نیست، می توانیم از روش های کلی تری، که به مدل سازی توام نشانگرهای گمشدگی و داده های طولی می پردازند، استفاده کنیم. تابع درستنمایی حاصل از این توزیع توام اصطلاحاً، تابع درستنمایی کامل نامیده می‌شود.(27)

لیتل در سال 1995 بیشتر این روش ها را در قالب دو کلاس از مدل ها تشریح می کند: مدل های گزینش و مدل های الگوهای آمیخته.(11)

لی و همکاران در سال 2007، چنین بیان می کنند که توزیع توام مذکور حداقل به سه شکل قابل تجزیه است:

1-تجزیه وابسته به پاسخ، به طوری که فرض می شود نشانگرهای پاسخ روی مقادیر اندازه های

تکراری شرطی شده‌اند.

2-تجزیه وابسته به الگو، به طوری که توزیع مقادیر اندازه های تکراری ترکیبی از توزیع های افراد درون زیرگروه های مشخص شده بر اساس الگوی گمشدگی است.

3-تجزیه وابسته به پارامتر، به طوری که مقادیر اندازه های تکراری و نشانگرهای پاسخ به شرط گروهی از پارامترهای مشترک از هم مستقلند.(14)

متعاقباً، سه گروه از مدل ها بر اساس هر تجزیه، تولید و نام گذاری می شوند: مدل های گزینش، مدل های الگوی آمیخته و مدل های پارامترهای مشترک. مدل پارامترهای مشترک در حقیقت می‌تواند در گروه مدل های گزینش نیز در نظر گرفته شود،(18). در ادامه به اختصار به معرفی این سه مدل می پردازیم. لازم به ذکر است که هر یک از این مدل ها می‌توانند به صورت حاشیه ای(برای مطالعه میانگین گروهی)، یا همراه با اثرات تصادفی(برای مطالعه ناهمگنی بین افراد) و یا به صورت انتقال(با شرطی شدن بر روی مشاهدات قبلی) ساخته شوند. علاوه بر این استفاده از این مدل ها در شرایط MCAR و MAR نیز بلامانع است؛ هر چند به دلیل پیچیدگی های محاسباتی، استفاده از آن ها تحت این شرایط توصیه نمی شود. در حقیقت در صورت برقرار نبودن فرض MNAR، این مدل ها به مدل های استاندارد تحلیل داده های طولی تبدیل می‌شوند؛ بنا بر این می توان آن ها را به عنوان مدل های مادر برای مدل های معمول حاشیه ای، اثرات تصادفی و انتقال در نظر گرفت.

پس از معرفی هر مدل، کاربرد آن در مثالی واقعی مورد بحث قرار می گیرد. داده های این مثال مربوط به کارآزمایی بالینی ترک اعتیاد مصرف کنندگان تنباکوی حوی متادون است. این مطالعه اثر دو درمان پیش‌گیری از بازگشت به اعتیاد(relapse prevention یاRP) و مدیریت وابستگی(contingency management یا CM) را به تنهایی و در ترکیب با هم را بر ترک اعتیاد مقایسه می کند. در کل 174 نفر در این مطالعه حضور داشته اند که به طور تصادفی در 4 گروه تیماری قرار گرفته اند: یک گروه کنترل که هیچ درمانی دریافت نکرده‌اند(42 نفر)؛ گروه RP(42 نفر)؛ گروه CM(43 نفر) و گروه RP+CM(47 نفر). هر فرد تحت 36 اندازه گیری مکرر در مدت 12 هفته، هر هفته 3 بار قرار گرفت. متغیر پاسخ ثبت شده برای هر فرد، میزان کربن مونوکسید موجود در بازدم او بود. گمشدگی در این داده ها هم به صورت یکنواخت و هم به صورت غیریکنواخت رخ داده بود. با توجه به مباحث قبلی، با فرض قابل چشم پوشی بودن، گمشده های متناوب با روش MCMC به تعداد چهار بار جانهی شدند(جانهی چندگانه). هر چهار مجموعه داده تولید شده بر اساس مدل های گزینش و پارامترهای مشترک تحلیل و در نهایت نتایج حاصل با قوانین مربوط به جانهی چندگانه ترکیب شدند.

از نظر تئوری، توزیع توام مشاهدات و الگوی گمشدگی بر اساس تابع درستنمایی زیر باید مدل سازی شوند:

$$L(\theta, \varphi|y_i^{obs}, X_i, r_i) \propto \prod\int f(y_i, r_i|X_i,\theta , \varphi)dy_i^{mis}$$

که در آن θ و φ به ترتیب پارامترهای مدل اندازه گیری و پارامترهای مکانیزم گمشدگی را نشان می‌دهند.

3-2-1 مدل گزینش

در مدل گزینش، تابع توزیع توام $f(y_i, r_i|X_i,\theta, \varphi)$ به تابع توزیع کناری y_i و تابع توزیع شرطی r_i به شرط y_i تجزیه می‌شود:

$$f(y_i, r_i|X_i,\theta , \varphi)= f(y_i |X_i, \theta) f(r_i| y_i, X_i, \varphi)$$

استفاده از مدل های گزینش در رویارویی با داده های گمشده در مطالعات طولی، تاریخچه نسبتاً طولانی دارد. این مدل اولین بار توسط هکمن در سال 1976 در متون اقتصادی معرفی شد که در آن، توزیع توام پاسخ دو متغیره y با گمشدگی در مولفه دوم، با استفاده از توزیع نرمال دو متغیره مشخص شد،(28). علاوه بر این، به طور ضمنی برای توزیع نشانگرهای پاسخ یک مدل خطی پروبیت در نظر گرفته شد. در حقیقت، مدل گزینش در فرمول بندی اولیه خودش شامل دو مرحله است. مرحله اول توسعه یک مدل پیش‌بینی با استفاده از متغیرهای پایه است به طوری که مشخص شود آیا یک فرد ریزش می‌کند یا نه. این مدل از ریزش، برای هر فرد یک احتمال ریزش مهیا می کند که این احتمال ها در مرحله دوم، تعیین مدل

طولی، به عنوان مخدوش‌گر وارد مدل می شوند تا مدل نسبت به اثر احتمالی ریزش، تعدیل شود. دیگل و کنوارد در سال 1994، این روش را با افزودن مقادیر گذشته متغیر وابسته و مقدار مشاهده نشده این متغیر در زمان ریزش، به مدل پیشگویی کننده ریزش گسترش دادند،(29). مدل آن ها گسترش مدل هکمن در مطالعات طولی با داده های ریزش شده بود که در آن از مدل لجستیک برای مخاطره ریزش استفاده شد. از آن پس مقالات زیادی در اجرا و گسترش مدل گزینش دیگل و کنوارد به چاپ رسید. از جمله تروکسل و همکاران این مدل ها را برای کنترل گمشدگی های غیریکنواخت گسترش دادند،(30). در سال 1995 نیز بیکر،(31)، و در همان سال فیتزمورس و همکاران،(32)، این مدل ها را برای متغیر پاسخ گسسته نیز توسعه دادند. علاوه بر این فعالیت هایی نیز در زمینه مطالعات بقا بر اساس مدل های گزینش انجام گرفت،(33-35). برای اطلاعات بیشتر در این زمینه، به کتاب فیتزموریسو همکاران مراجعه کنید،(6). برای درک بهتر موضوع به کاربرد مدل انتخاب در داده های کارآزمایی بالینی ترک سیگار می پردازیم.

مدل انتخاب در نظر گرفته شده برای تحلیل این داده ها به صورت زیر است:

برای اندازه های تکراری، یک مدل خطی اثرات تصادفی با ساختار کواریانس AR(1) به صورت زیر در نظر گرفته شد:

$$y_{ij}=\beta_0+\beta_1 CM_i+\beta_2 RP_i+\beta_3(CM_i\times RP_i)+\beta_4 baseCO_i+\beta_5 patches_i$$

به طوری که CM_i و RP_i به ترتیب نشانگرهای مربوط به گروه تیماری CM و RP برای فرد i ام هستند؛ $baseCO_i$ میزان اولیه مونوکسیدکربن اندازه گیری شده برای فرد و $patches_i$ تعداد بسته های نیکوتینی است که شخص در طول مطالعه دریافت کرده است. مدل انصراف استفاده شده نیز مدل رگرسیون لجستیک به صورت زیر بود:

$$Logit(p_{di}(y_{i,di}, H_{ij}))=¢_0+¢_1 y_{i,di}+¢_2 y_{i,di}-1$$

به طوری که d_i، زمان انصراف را برای فرد i ام نشان می دهد. نتایج نهایی حاصل از برازش این مدل انتخاب در جدول شماره 2 آورده شده است:

جدول شماره 2. برآورد اثرات تیمار و پارامترهای مدل انصراف در مدل گزینش

$¢_1$(SD)	$¢_1$(SD)	β_3(SD)	β_2(SD)	(SD)β_1
-۰،۰۳	۱،۲۸	-۰،۰۸	۰،۰۲	-۰،۲۸
(۰،۲۳)	(۰،۳۳)	(۰،۰۷)	(۰،۰۵)	(۰،۰۵)

در مدل فوق تنها اثر درمان CM معنی‌دار بود. اثر درمان RP و اثر متقابل بین CM و RP معنی دار نشد. ضریب رگرسیونی $¢_1$ نیز به طور معنی داری بزرگ تر از صفر است($¢_1$ =۱،۲۸ و P=0.0002) که بیانگر این مطلب است که هر چه داده گمشده موردنظر بزرگ تر باشد، احتمال انصراف بیشتر خواهد بود. به عبارت دیگر انصراف احتمالاً غیرقابل چشم پوشی است.

3-2-2-مدل الگوی آمیخته

در مدل الگوهای آمیخته که یک مدل وابسته به الگو است، فرض می‌شود که توزیع اندازه های مکرر بسته به الگوی گمشدگی تغییر می‌کند، بنا بر این تابع توزیع توام فوق به صورت زیر تجزیه می شود:

$$f(y_i, r_i|X_i,\theta , \varphi)= f(y_i \mid r_i , X_i, \theta)\, f(r_i|X_i, \varphi)$$

گسترش مدل های گزینش برای کنترل گمشده های آگاهی بخش به سال 1977 بر می گردد؛ در آن سال روبین روش هایی را برای استفاده از پیشین های آگاهی بخش در جهت بهره برداری از اطلاعات افراد بی پاسخ بررسی کرد،(36). لیتل در سال های 1993 و 1994 به عنوان راهکار دیگری در مقابل مدل های گزینش، گروهی کلی از مدل ها را تحت عنوان«مدل‌های الگوی آمیخته» برای تحلیل داده های گمشده معرفی و فرمول بندی کرد،(۳۷،۳۸). در این زمینه قبل از لیتل هم فعالیت هایی صورت گرفته بود؛ از جمله توسط مارینی و همکاران،(39)، در سال 1980 و گلین و همکاران،(40)، در سال 1986.

ولی مقاله های لیتل این مدل ها را به طور کامل و به صورت آماری به روشی کلی معرفی کرند. پس از آن مقاله های زیادی در معرفی و گسترش این مدل ها برای داده های طولی به چاپ رسید،(43-41). برای اطلاعات بیشتر به کتاب دانیالز و هگان رجوع کنید.(11)

در این مدل ها، افراد بر اساس الگوی گمشدگی به گروه های مختلف تقسیم می شوند. این گروه ها در تحلیل های بعدی می توانند به عنوان مثال، برای بررسی اثر الگوی گمشدگی روی متغیر پاسخ مورد نظر استفاده شوند. در استفاده از روش الگوی آمیخته، مدلی مشخص می شود که نیازی به فرض قابل چشم پوشی بودن گمشدگی ندارد.(11)

در داده های کارآزمایی بالینی ترک سیگار، به دلیل تعداد زیاد اندازه گیری ها، تعداد الگوهای گمشدگی ممکن موجود در داده ها زیاد است. در این جا برای سادگی، الگوهای مشاهده شده به دو گروه داده های کامل و داده های ریزش شده تقسیم شده اند. یک مدل خطی با اثرات تصادفی با ساختار کواریانس AR(1) برای تحلیل میزان مونوکسیدکربن از هفته دوم به بعد به صورت زیر در نظر گرفته شد:

$$y_{ij}=\beta_0+\beta_1 CM_i+\beta_2 baseCO_i+\beta_3 patches_i$$

مدل فوق به طور جداگانه برای افراد با داده های کامل و افراد با داده های ریزش شده، برازش داده شد. برآورد به دست آمده برای درمان CM، برابر با β_1 با انحراف استاندارد 0/13 بود. آزمون بر اساس آماره t منجر به P=0.06 شد.

3-2-3-مدل گزینش در مقابل مدل الگوی مشترک

هر یک از مدل های معرفی شده گزینش و الگوهای مشترک مزایا و معایب خاص خود را دارد. در حالت کلی تمام روش های مبتنی بر درستنمایی برای کنترل گمشده های غیرقابل چشم‌پوشی، بر پایه فرضیاتی غیرقابل تحقیق استوارند؛ زیرا مکانیزم داده های گمشده(MNAR) به داده های مشاهده نشده بستگی دارند. مسلماً صحت و اعتبار این مدل ها به برقراری یا عدم برقراری این فرضیات وابسته اند. علاوه بر توجه به این نکته، در استفاده از این مدل ها باید در زمینه های اریبی و کارایی برآوردها، حساسیت برآوردها نسبت به مدل انتخابی، پیچیدگی های محاسباتی و قابلیت اجرا و تفسیر مدل و نتایج آن دقت کرد.(۶،۷،۱۸)

در اکثر مطالعات هدف اصلی، تعیین ارتباط متغیر پاسخ با متغیرهای مستقل است. از آن جا که مدل گزینش به طور مستقیم این ارتباط(توزیع حاشیه ای پاسخ های طولی) را مدل سازی می‌کند، استفاده از آن معمولاً مفیدتر و قابل فهم‌تر است. علاوه بر این در مدل های انتخاب با وجود مدل شرطی مکانیزم گمشدگی داده ها، انجام آزمون فرضیات درباره این مکانیزم ساده خواهد بود. البته در استفاده از این مدل ها با محدودیت هایی نیز مواجه هستیم. مشکل عمده مدل های گزینش، حساسیت زیاد برآوردهای حاصل از آن به مدل انتخاب شده است. علاوه بر این محاسبات لازم جهت برازش این مدل معمولاً پیچیده است.(۷،۱۸)

در مدل های الگوهای مشترک، فرضیات خاصی درباره توزیع پاسخ های مشاهده نشده در نظر گرفته می شود که حساسیت نتایج نسبت به مدل انتخابی را کاهش می دهد. بنا بر این این مدل ها از این جهت نسبت به مدل های گزینش برتری دارند ولی مشکل اصلی آن ها، این است که پارامترهای مورد نظر محقق (در ارتباط متغیرهای وابسته و مسستقل)، مستقیماً در دسترس نیستند. استنباط‌های اولیه در این مدل ها، بر اساس توزیع های پاسخ به شرط الگوی گمشدگی است. به عبارتی در هر الگوی گمشدگی، مدل سازی به صورت مجزا انجام و پارامترها برآورد می شوند. بنا بر این استنباط در مورد توزیع حاشیه ای پاسخ ها که هدف اصلی بسیاری از مطالعات است، با ترکیب نتایج به دست آمده در مرحله اول ممکن است. در این صورت، بررسی اثرات تصادفی فردی روی توزیع حاشیه ای پاسخ ها در قالب ضرایب رگرسیونی غیر ممکن است. به علاوه برازش مدل های الگوی مشترک نیز خالی از پیچیدگی های محاسباتی نیست.(۶،۷)

3-2-4-مدل پارامترهای مشترک

مدل پارامترهای مشترک، فرض می کند y_i و r_i استقلال شرطی دارند(به شرط گروهی از پارامترهای

مشترک) و در نتیجه تابع توزیع توام به صورت تجزیه می شود.(16)

$$f(y_i, r_i|X_i,\theta , \varphi)=\int f(y_i \mid v_i , X_i, \theta) f(r_i| v_i , X_i, \varphi)dv_i$$

«پارامترهای» مشترک در این مدل ها نقش مخدوش کننده را در ارتباط بین y_i و r_i بازی می کنند؛ بنا بر این می توانند شامل عوامل قابل مشاهده(مانند جنسیت) یا پنهان(یعنی اثرات تصادفی) باشند. این مدل ها برای اولین بار توسط وو و کارول،(44) در سال 1988 معرفی شدند. آن ها در مطالعه ای طولی که به بررسی عملکرد ریه پرداخته بودند برای متغیر پاسخ یک مدل خطی ساده با شیب و عرض از مبدا تصادفی و برای مکانیزم گمشدگی یک مدل پروبیت در نظر گرفته بودند که شیب تصادفی به عنوان پیشگو در آن وارد شده بود. وقتی ضریب رگرسیون پروبیت برای این شیب تصادفی مخالف صفر باشد، بین پاسخ و فرایند گمشدگی وابستگی وجود دارد. عدم توجه به این وابستگی باعث ایجاد اریبی در برآوردها می شود. بنا بر این چنان چه بیان شد، مدل های پارامترهای مشترک نیز روشی برای غلبه بر گمشدگی های غیرقابل چشم پوشی و یا آگاهی بخش خواهند بود. مقاله های کلیدی در معرفی و گسترش این مدل ها شامل وو و کارول در سال 1988، فلمن و وو،(45)، در سال 1995، وولفسن و همکاران،(46)، در سال 1997 هستند.

بحث و نتیجه گیری

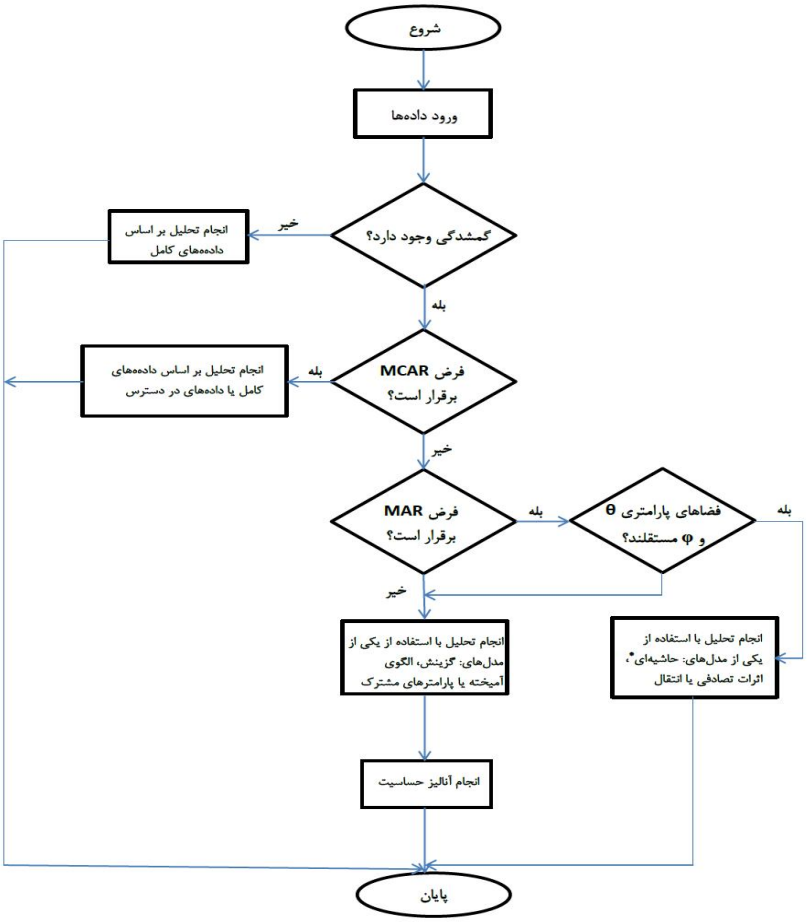
مشکل وجود داده های گمشده در مطالعات طولی و کارآزمایی های بالینی، در سال های اخیر به شدت مورد توجه قرار گرفته است، به طوری که متون زیادی در این راستا نگارش و به چاپ رسیده است. در سال 2000 وربک و مولنبرگز،(47)، 2004 فیتزموریس و همکاران،(18)، 2005 مولنبرگز و وربک،(48)، 2007 مولنبرگز و کنوارد،(49)، 2008 دانیالز و هگان،(11)، و در همان سال فیتزموریس و همکاران،(6)، و در سال 2009 ابراهیم و مولنبرگز،(7)، به طور گسترده به مفاهیم گمشدگی و مدل های مناسب برای آن ها پرداختند.

با توجه به اهمیت موضوع در مقاله حاضر به صورت کاربردی به بیان مفاهیم و مدل های آماری موجود در این زمینه پرداخته شد. گمشدگی و پیامدهای ناشی از آن، الگوها و مکانیزم های گمشدگی، قابل چشم پوشی بودن و مدل های مناسب برای آن و در نهایت مدل های مناسب برای گمشده های غیرقابل چشم پوشی موضوع های مباحث مطرح شده در مقاله حاضر را در بر می‌گیرند.

در سال های اخیر علاوه بر مباحث تئوری، پیشرفت های نرم افزاری چشم گیری نیز در راستای مدل سازی داده های طولی با حضور گمشدگی، صورت گرفته است. اکثر نرم افزارهای آماری از جمله SAS، S-plus، R، SPSS و Stata قابلیت اجرای مدل های مربوط به گمشده های قابل چشم پوشی معرفی شده در این مقاله را دارند،(18). در مقابل در مورد گمشدگی های غیرقابل چشم پوشی معمولاً نیازمند به کدنویسی های پیچیده هستیم. تاکنون تلاش هایی در این زمینه در نرم افزارهای SAS، R و WinBUGS انجام شده است.(11)

با توجه به این‌که مقادیر گمشده همواره مشاهده نشده باقی خواهند ماند، تمام روش های موجود برای تحلیل داده های گمشده شامل فرضیات غیرقابل شناسایی و غیرقابل اثبات هستند. بنا بر این بهتر است در برخورد با این داده ها صرفاً به نتایج حاصل از یک روش اتکا نشود و در این راستا تحلیل های حساسیت مناسب نیز انجام شود.

به طور خلاصه طبق مباحث مطرح شده در این مقاله، می توان الگوریتم زیر را برای استنباط‌های مبتنی بر مدل استفاده کرد:



شکل شماره 1. فلوچارت بررسی گمشدگی

سپاسگزاری

مقاله حاضر بخشی از پایان نامه کارشناسی ارشد آمار زیستی در دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی است. بدین وسیله از کلیه همکاران دانشکده پیراپزشکی تشکر و قدردانی می نماییم.

References

1-Rubin DB. Multiple Imputation for Non-response in Surveys. New York: John Wiley & Sons; 1987.
2-Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.
3-Molenberghs G, Kenward MG, Goetghebeur E. Sensitivity analysis for incomeplete contingency tables: the Slovenian plebiscite case. J Appl Statist 2001;50:15-29.
4-Robins JM, RotnitzkyA. Recovery of information and adjustment for dependent censoring using surrogate markers. Boston: Birkhauser; 1992.
5-Scharfstein DO, Robins JM. Estimation of the failure time distribution in the presence of informative censoring. Biometrika 2002;89:617-34.
6-Fitzmaurice GM, Davidian M, Verbeke G, Molenberghs M. Longitudinal data analysis. London: Chapman & Hall; 2008.
7-Ibrahim JG, Molenberghs M. Missing data methods in longitudinal studies: a review. NIH 2009;18:1-43.
8-Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13-22.
9-Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963-74.

10-Diggle P, Heagerty P, Liang KY, Zeger S. Analysis of Longitudinal Data. 2nd ed. Oxford: Oxford University Press; 2002.
11-Daniels MJ, Hogan JW. Missing data in longitudinal studies. London: Chapman & Hall; 2008.
12-Li J, Yang X, Wu Y, Shoptaw S. A random-effects Markov transition model for Poisson-distributed repeated measures with nonignorable missing values.Stat Med 2007;26:2519-32.
13-Shoptaw S, Rotheram FE, Yang X, Frosch D, Nahom D, Jarvik ME, et al. Smoking cessation in methadone maintenance. Addiction 2002;97:1317-28.
14-Diggle P, Heagerty P, Liang KY, Zeger S. Analysis of Longitudinal Data. 2nd ed. Oxford University Press: Oxford; 2002.
15-Twisk J. Applied Longitudinal Data Analysis for Epidemiology. New York: Cambridge University Press; 2003.
16-Yang X, Li J, Shoptow S. Imputation-based strategies for clinical trial longitudinal data with nonignorable missing values. Stat Med 2008;27:2826-49.
17-Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nded.New York: Wiley; 2002.
18-Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. New York: Wiley; 2004.
19-Nelder JA, Wedderburn RWM. Generalized linear model.J Roy Statist Soc Ser A 1972; 135:370-84.
20-Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. Statist Med 1998; 17:1643-58.
21-Casella G, Berger. Statistical inference. Duxbury Press; 2002.
22-Lipsitz SR, Fitzmaurice JG, Ibrahim M, Sinha DM, ParzenLipshultz S. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. J Roy Statist Soc Ser A 2009;1:3-20.
23-McCullagh P. Regression models for ordinal data (with discussion). J Roy Statist Soc Ser B 1980;42:109-42.
24-Myers RH, Montgomery DC, Vining GG. Generalized linear models. USA: Wiley & Sons; 2002.
25-Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Amer Statist Assoc 1995; 90:106-21.
26-Hedeker D, Gibbons RD. Longitudinal data analysis. New Jersey: Wiley; 2008.
27-Schafer JL. Analysis of Incomplete Multivariate Data. London: Chapman & Hall; 1997.
28-Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models.Ann Econ Soc Meas 1976;5:475-92.
29-Diggle PJ, Kenward MG. Informative drop-out in longitudinal data analysis. J Appl Statist 1994;43:49-93.
30-Troxel AB, Harrington DP, Lipsitz SR. Analysis of longitudinal data with non-ignorable non-monotone missing values. J Appl Statist 1998; 47:425-38.
31-Baker SG. Marginal regression for repeated binary data with outcome subject to nonignorable non-response. Biometrics 1995;51:1042-52.
32-Fitzmaurice GM, Molenberghs G, Lipsitz SR. Regression models for longitudinal binary responses with informative dropouts. JRoy Statist Soc Ser B 1995;57:691-704.
33-Schluchter MD. Methods for the analysis of informatively censored longitudinal data. Stat Med 1992;11:1861-70.
34-DeGruttola V, Tu XM. Modeling progression of CD4 lymphocyte count and its relationship to survival time. Biometrics 1994;50:1003-14.
35-Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. J Amer Statist Assoc 1995;90:27–37.
36-Rubin DB. Formalizing subjective notions about the effect of non-respondents in sample surveys.J Amer Statist Assoc 1977; 72:538-43.
37-Little RJA. Pattern-mixture models for multivariate incomplete data. J Amer Statist Assoc 1993;88:125-34.
38-Little RJA. A class of pattern-mixture models for normal incomplete data. Biometrika 1994;81:471-83.
39-Marini M, Olsen AR, Rubin DB. Maximum-likelihoodestimation in panel studies with attrition. SM 1980; 314-57.

40-Glynn RJ, Laird NM, Rubin DB. Selection modeling versus mixture modeling with non-ignorable nonresponse. In:Wainer H, editor. Drawing Inferences from Self Selected Samples. New York: Springer; 1986.P.115-42.
41-Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D. Strategies to fit pattern-mixture models. Biostatistics 2002; 3:245-65.
42-Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable dropout. Stat Med 2003;22:2553-75.
43-Demirtas H. Multiple imputation under Bayesian smoothed pattern-mixture models for non-ignorabledrop-out. Stat Med 2005; 24:2345-63.
44-Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics 1988;44:175-88.
45-Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. Biometrics 1995;51:151-68.
46-Wulfsohn MS,Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics 1997;53:330-9.
47-Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for non-random dropout: a local influence approach. Biometrics 2001;57:7-14.
48-Molenberghs G; Verbeke G. Models for discrete longitudinal data. New York: Springer; 2005.
49-Molenberghs G, Kenward MG. Missing data in clinical studies. New York: Wiley; 2007.

Different Types of Missing in Longitudinal Data and the Likelihood-base Methods Applied in their Analysis

ZayeriF[1] , Akbarzadeh Baghban A.R[2], Kazemzadeh M[3], Yaseri M[4], Abbasi A.M[5]*

(Received: 7 Nov. 2012 Accepted: 11 Feb. 2013)

Abstract

Missing values are frequently seen in data sets of research studies conducted in different sciences such as medicine and especially in longitudinal studies in which every individual are exposed to the repeated measures over time. In the last few decades, a vast majority of statistical activities has been done in this area, including the areas of concepts, issues, and theoretical and software methods. Despite the widespread use of the results of these statistical activities, the researchers, in many cases, have been seen to have a vague impression from these concepts which results in inaccurate inferences. Therefore, given the importance of the issue and the need for the scientific community to know these issues correctly and accurately, the current study is set to review and compare the concepts such as missing data patterns and mechanism, as well as the existing models in analyzing longitudinal data with missing values. Furthermore, their application will be explored in the data obtained from a clinical trial of addiction treatment with a continuous response variable.

Keywords: missing completely at random, missing at random, missing not at random, selection model, pattern-mixture model, shared parameters model

1. Proteomics Research Center, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
2. Dept of Basic Sciences, Faculty of Rehabilitation, Shahid Beheshti University of Medical Sciences, Tehran, Iran
3. Dept of Biostatistics, Faculty of Paramedical Sciences Shahid Beheshti University of Medical Sciences, Tehran, Iran
4. Dept of Epidemiology and Biostatistics, Faculty of Health, Tehran University of Medical Sciences, Tehran, Iran
5. Dept of Occupational Health, Faculty of Health, Ilam, Iran
**(corresponding author)*