

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

فرید زایری^{1*}، علی رضا اکبرزاده باغبان²، مژگان کاظم زاده³، مهدی یاسری⁴، علی محمد عباسی⁵

- (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 آمار زیستی، 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 می گیرد، 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.

واژه های 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 می گیرد. 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 می آورد.

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 ساده ترین 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 بوده اند (complete-case analysis). متأسفانه 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 می شوند. استراتژی های دیگر شامل روش های جانهی (imputation)، (3-1)، روش های وزن دهی، (5، 4)، 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 می شود (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 می دهد. 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 می شود 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 می شوند. 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 کرده اند: 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 می شوند. 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

مدل‌ها به تحلیل داده‌های 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 کارآزمایی‌های بالینی در علوم مختلف از جمله علوم پزشکی بسیار گسترده است. مشکل عمده استفاده از این طرح‌ها، وجود داده‌های گمشده است. هرگاه یک یا چند دنباله از اندازه‌های مکرر مربوط به واحدها یا افراد درون مطالعه ناکامل باشد، با مقادیر گمشده در مجموعه داده‌ها مواجه هستیم، (14)، (هر چند مفهوم گمشدن ساده به نظر می‌رسد ولی در مواردی در نظر گرفتن داده‌های ثبت نشده به عنوان گمشده بی‌معنا خواهد بود. توجه به این نکته ضروری است که منظور از داده گمشده مقداری است که بر اساس طرح مطالعه باید ثبت می‌شده ولی به دلایلی ثبت نشده است. بنا بر این چنانچه طراحی مطالعه به گونه‌ای باشد که برای بعضی از افراد زمان‌های بیشتری برای ثبت اندازه‌های مکرر و برای بعضی دیگر زمان‌های 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 داده‌های ثبت نشده افراد فوت شده در حین مطالعه، به عنوان داده‌های گمشده بی‌معنا است، (15). حضور این مقادیر دلایل مختلفی می‌تواند داشته باشد از جمله در دسترس نبودن افراد، کناره‌گیری افراد از مطالعه، بروز اثرات جانبی منفی، موثر نبودن مداخله، شرایط سخت درمان و ...، (16). به طور کلی حضور مقادیر گمشده در داده‌ها سه مشکل عمده ایجاد می‌کنند: نخست این که قسمتی از اطلاعات از دست می‌رود، 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 داده‌ای نامتعادل می‌شود، زیرا همه افراد دارای تعداد یکسانی از اندازه‌گیری‌های مکرر در زمان‌های یکسان نخواهند بود. بنا بر این امکان استفاده مستقیم از مدل‌ها و روش‌هایی که برای داده‌های کامل قابل اجرا هستند، وجود ندارد. علاوه بر موارد فوق، از آنجا که ممکن است اطلاعات گمشده مربوط به افرادی باشد که

تفاوت‌های اساسی با افراد دیگر (افراد با اطلاعات کامل) داشته باشند، پتانسیل آریبی وجود دارد. برطرف کردن آریبی در این موارد مشکل است زیرا دلایل دقیق گمشدگی معمولاً ناشناخته‌اند. (۱۶)

گمشدگی در مطالعات طولی به طور کلی می‌تواند 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 داده‌های گمشده
(missing data mechanisms)

بحث گمشدگی در مطالعات طولی حساس‌تر از مطالعات مقطعی است زیرا گمشدگی می‌تواند در اشکال و موقعیت‌های مختلفی رخ دهد، (1). به همین دلیل در تحلیل داده‌های طولی با گمشدگی، معمولاً نیاز به فرضیاتی درباره دلایل گمشدگی داریم که مکانیزم داده‌های گمشده نامیده می‌شوند. در حالت کلی وقتی می‌توانیم استنباط‌های معتبر در مورد داده‌های ناکامل داشته باشیم که مکانیزم داده‌های گمشده به طور کامل شناخته شده باشند ولی از آنجایی که این مکانیزم تحت کنترل محقق نیست، شناخت دقیق آن امکان‌پذیر نیست. بنا بر این در هر مطالعه فرضیاتی درباره این مکانیزم پذیرفته می‌شود. 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 سوال‌های کاربردی تحقیق دارند یا نه؟» شکل می‌گیرند. چنانچه دلایل گمشدگی با پیامد مورد نظر مرتبط نباشند، مشکل خاصی در تحلیل داده‌ها وجود نخواهد داشت. ولی در صورتی که این ارتباط وجود داشته باشد، ممکن است منجر به برآوردهای آریب شود؛ زیرا داده‌های در دسترس با داده‌های کامل متفاوت خواهند بود. مسلماً اعتبار تحلیل‌ها به برقراری یا عدم برقراری این فرضیات وابسته است. (۱۴، ۶)

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

شده وابسته باشد ولی به پاسخ‌های مشاهده نشده بستگی نداشته باشد. یعنی:

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

چنانچه مشاهده می‌شود در این نوع از گمشدگی، احتمال گم شدن ممکن است به مشاهدات قبلی وابسته باشد. این مکانیزم نسبت به مکانیزم MCAR معقول‌تر به نظر می‌رسد، هر چند باعث ایجاد محدودیت‌های بیشتر در تحلیل خواهد شد، زیرا در مکانیزم MAR پاسخ‌های مشاهده شده، نمونه‌ای تصادفی از کل داده‌ها نخواهد بود.

گمشده‌های ساختاری (missing by design) نمونه‌ای از گمشده‌های 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 غیر تصادفی وقتی رخ می‌دهد که احتمال مشاهده نشدن پاسخ در یک زمان، حداقل به یکی از پاسخ‌هایی که باید مشاهده می‌شده‌اند ولی گم شده‌اند 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 مطالعه‌ای را در نظر بگیرید که هدف آن اندازه‌گیری کیفیت زندگی افراد است (متغیر پاسخ). چنانچه افرادی که سطح کیفیت زندگی بالا یا پایین برخوردارند، از پرسش‌نامه و کامل کردن مطالعه اجتناب کنند، گمشدگی غیر تصادفی رخ داده است. به عنوان مثالی دیگر مطالعه‌ای را در نظر بگیرید که بر روی بیماران تنفسی در شهری خاص صورت می‌گیرد و علت عمده گمشدگی داده‌های آن، مهاجرت بیماران باشد. چنانچه 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.

ناقص ارائه دادند که شامل دسته بندی مفیدی از مکانیزم‌های 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 $y_i = (y_{i1}, \dots, y_{ij})$ خواهد بود. برای هر بردار پاسخ برداری از متغیرهای نشانگر به صورت $r_i = (r_{i1}, \dots, r_{ij})$ 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 هیچ‌یک از پاسخ‌های 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 شرکت‌کننده در مطالعه در مناطق مختلف شهری ساکن هستند. چنانچه دلیل عدم ثبت داده‌های افراد، مسائلی چون بارش برف یا ترافیک سنگین در روز تعیین شده برای مراجعه به مرکز جهت اندازه‌گیری باشد، گمشدگی کاملاً تصادفی رخ داده است. زیرا گم شدن داده‌ها ارتباطی با مشاهدات قبلی بردار y_i ندارد. در حقیقت در مکانیزم MCAR، داده‌های مشاهده شده نمونه‌ای تصادفی از کل داده‌ها خواهند بود.

2- گمشدگی تصادفی (MAR)

گمشدگی تصادفی زمانی اتفاق می‌افتد که احتمال گم شدن پاسخ در یک زمان، به پاسخ‌های مشاهده

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 (X_i, y_i^{obs})). ب) پارامترهای فرایند اندازه گیری (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 جداگانه دو عبارت سمت راست است، (6)، چون عبارت

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

شامل هیچ اطلاعاتی 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 می شود. علاوه بر این در چارچوب استنباط های فراوانی گرا، تنها مکانیزم MCAR قابل چشم پوشی خواهد بود. (6)

2-3- 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 می شود 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 $y_{i, k+1}, \dots, y_{i, n}$ 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 اندازه گیری مراجعه نکرده ولی در دوره های بعدی در مطالعه حضور داشته است.

با وجود این که الگوهای متناوب می توانند 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 اندازه گیری 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. (1، 16)

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 طبقه بندی شوند (یعنی بر حسب y_i^{obs}) می‌توان 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 می‌دهد 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)

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 می‌تواند 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

در نهایت وقتی گمشدگی NMAR است، تقریباً هیچ یک از روش‌های استاندارد تحلیل داده‌های طولی معتبر نیستند. هم روش‌های GEE و هم روش‌های بر مبنای درست‌نمایی به برآوردهای اریب از میانگین پاسخ منجر می‌شوند، (18). بنا بر این روش‌های مدل‌سازی پیشرفته برای گمشده‌های غیرقابل چشم‌پوشی مورد نیاز است.

قبل از ورود به بحث گمشده‌های آگاهی بخش، به ارائه مثالی شبیه‌سازی شده در راستای مفاهیم فوق می‌پردازیم:

2-3- شبیه‌سازی MCAR، MAR و MNAR

هدگر و گینز در سال 2008 به مطالعه‌ای شبیه‌سازی شده جهت تشخیص روش‌های مناسب کنترل داده‌های گمشده در مطالعات طولی پرداختند، (26). داده‌ها بر اساس مدل زیر شبیه‌سازی شدند:

$$y_{ij} = \beta_0 + \beta_1 \text{time}_j + \beta_2 \text{group}_i + \beta_3 (\text{group}_i \times \text{time}_j) + v_{0i} + v_{1i} \text{time}_j + \varepsilon_{ij}$$

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 واریانس‌های $\sigma_{v_0}^2 = 4$ ، $\sigma_{v_1}^2 = 25$ و کواریانس $\sigma_{v_0 v_1} = -1$ و خطاهای ε_i دارای توزیع نرمال با میانگین صفر و واریانس $\sigma^2 = 4$ فرض شدند. ابتدا 5000 داده با پنج زمان اندازه‌گیری بر اساس مدل و پارامترهای ذکر شده تولید شد. سپس به تولید داده‌های گمشده با مکانیزم‌های MCAR، MAR و MNAR پرداختند. در نهایت برای تحلیل داده‌ها مدل اثرات تصادفی با عرض از مبدا تصادفی و روند زمانی و نیز مدل حاشیه‌ای با برآوردهای GEE استاندارد برازش داده شد. نتایج حاصل در جدول شماره 1 آمده است.

وقتی گمشده‌ها MAR هستند (و البته MCAR نیستند) روش‌های بر مبنای بردارهای کامل و یا روش GEE استاندارد، برآوردهای اریب از میانگین می‌دهند. در مقابل روش‌های بر مبنای درست‌نمایی که به طور دقیق توزیع توأم کامل پاسخ‌ها را مشخص می‌کنند، برآوردهای معتبر می‌دهند. البته یک شرط ظریف ولی مهم وجود دارد و آن این است که توزیع توأم 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، روش‌های بر مبنای درست‌نمایی، به شرط این که ماتریس کواریانس به درستی مدل شود، استنباط‌های معتبر به دست می‌دهند. به طور مشابه مدل‌های 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 توأم بردار پاسخ، بسیار حساس است. بنا بر این وقتی داده‌های طولی ناقص هستند در تعیین ارتباط درون فردی باید دقت کرد. (۱۸، ۱۱)

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 داده‌های مشاهده شده برقرار نیست، بنا بر این اعتبار تحلیل‌ها به خطر می‌افتد. با استفاده از یک GEE وزنی ساده، (۲۵، ۲۲)، روش‌هایی برای تعدیل تحلیل‌ها ارائه می‌شود که در آن وزن‌ها با استفاده از یک مدل برای احتمال گمشدگی برآورد می‌شوند. 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)

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

σ^2	σ_{v1}^2	σ_{v01}	σ_{v0}^2	B_3	β_2	B_1	β_0	مدل	
4	۰٫۲۵	-۰٫۱	4	-1	0	-1	25	---	مقادیر شبیه سازی
۴٫۰۷۰	۰٫۱۹۹	-۰٫۰۲۰	۳٫۹۱۸	-۰٫۹۸۶	-۰٫۰۰۱	-۰٫۹۹۴	۲۴٫۹۶۹	اثرات تصادفی	داده های کامل
(۰٫۰۴۶)	(۰٫۰۱۴)	(۰٫۰۳۲)	(۰٫۱۲۹)	(۰٫۰۲۳)	(۰٫۰۷۱)	(۰٫۰۱۶)	(۰٫۰۵۰)		
۴٫۰۷۰	۰٫۱۹۹	-۰٫۰۲۰	۳٫۸۱۱	-۰٫۹۳۳	-۰٫۰۸۷	-۱٫۰۲۴	۲۴٫۹۹۱	اثرات تصادفی	MCAR
(۰٫۰۸۳)	(۰٫۰۲۵)	(۰٫۰۵۶)	(۰٫۱۹۳)	(۰٫۰۳۲)	(۰٫۰۸۹)	(۰٫۰۲۳)	(۰٫۰۶۳)		
3٫۸۷۳	۰٫۲۳۳	-۰٫۰۶۴	۳٫۹۸۱	-۰٫۹۶۹	-۰٫۰۱۰	-۱٫۰۳۹	۲۴٫۹۹۶	اثرات تصادفی	MAR
(۰٫۰۷۸)	(۰٫۰۳۲)	(۰٫۰۶۵)	(۰٫۱۵۸)	(۰٫۰۴۱)	(۰٫۰۷۵)	(۰٫۰۲۵)	(۰٫۰۵۳)		
				-۱٫۰۰۱	۰٫۰۱۹	-۱٫۱۶۴	25٫281	حاشیه ای	
				(۰٫۰۸۵)	(۰٫۰۹۷)	(۰٫۰۳۷)	(۰٫۰۵۸)		
۳٫۰۲۰	۰٫۳۱۹	-۰٫۰۴۳	۳٫۸۵۶	-۰٫۵۵۲	۰٫۰۲۷	-۰٫۲۳۳	۲۴٫۹۵۶	اثرات تصادفی	MNAR
(۰٫۰۵۳)	(۰٫۰۲۵)	(۰٫۰۵۱)	(۰٫۱۳۱)	(۰٫۰۳۵)	(۰٫۰۷۰)	(۰٫۰۲۰)	(۰٫۰۴۹)		
				-۰٫۵۸۳	۰٫۰۱۶	-۰٫۳۸۶	۲۵٫۰۵۱	حاشیه ای	
				(۰٫۰۳۴)	(۰٫۰۷۱)	(۰٫۰۲۰)	(۰٫۰۴۹)		

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 می شود. (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

با توجه به جدول شماره 1 مشاهده می شود که در حالت MCAR، برآوردهای به دست آمده از برازش مدل اثرات تصادفی به خوبی مقادیر واقعی پارامترها را پوشش می دهند، 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 می دهد. 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 می دهد که در حالت MAR مدل اثرات تصادفی برآوردهایی معقول از پارامترها به دست می دهد، در حالی که برازش مدل حاشیه ای منجر به چنین برآوردهایی نمی شود. تولومی و همکاران نیز در مطالعه ای مشابه نشان دادند که اریبی ناشی از برازش مدل حاشیه ای با GEE استاندارد، با افزایش درصد و شدت غیرتصادفی بودن گمشدگی، افزایش می یابد.

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 نبوده اند. بنا بر این در شرایط MNAR، استفاده از مدل های تحلیل داده های MCAR و MAR مشکل ساز خواهد بود.

تکراری شرطی شده‌اند.

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

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

متعاقباً، سه گروه از مدل‌ها بر اساس هر تجزیه، تولید و نام گذاری می‌شوند: مدل‌های گزینش، مدل‌های الگوی آمیخته و مدل‌های پارامترهای مشترک. مدل پارامترهای مشترک در حقیقت می‌تواند در گروه مدل‌های گزینش نیز در نظر گرفته شود، (18).

در ادامه به اختصار به معرفی این سه مدل می‌پردازیم. لازم به ذکر است که هر یک از این مدل‌ها می‌توانند

به صورت حاشیه‌ای (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 مدل‌ها در شرایط MCAR و MAR نیز بلامانع است؛ هر چند به دلیل پیچیدگی‌های محاسباتی، استفاده از آن‌ها تحت این شرایط توصیه نمی‌شود. در حقیقت در صورت برقرار نبودن فرض MNAR، این مدل‌ها به مدل‌های استاندارد تحلیل داده‌های طولی تبدیل می‌شوند؛ بنا بر این می‌توان آن‌ها را به عنوان مدل‌های مادر برای مدل‌های معمول حاشیه‌ای، اثرات تصادفی و انتقال در نظر گرفت.

پس از معرفی هر مدل، کاربرد آن در مثالی واقعی مورد بحث قرار می‌گیرد. داده‌های این مثال مربوط به کارآزمایی بالینی ترک اعتیاد مصرف کنندگان تنباکوی حوی متادون است. این مطالعه اثر دو درمان پیش‌گیری از بازگشت به اعتیاد (RPl relapse prevention) و مدیریت وابستگی (contingency management) یا (CM) را به تنهایی و در ترکیب با هم را بر ترک اعتیاد مقایسه می‌کند. در کل 174 نفر در این مطالعه حضور داشته‌اند که به طور تصادفی در 4 گروه تیماری قرار گرفته‌اند: 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 اندازه‌گیری 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 داده‌ها هم به صورت یکنواخت و هم به صورت غیریکنواخت رخ داده بود. با توجه به مباحث قبلی، با فرض قابل چشم‌پوشی بودن، گمشده‌های متناوب با روش MCMC به تعداد چهار بار جانمایی شدند (جانمایی چندگانه). هر چهار مجموعه داده تولید شده بر اساس مدل‌های گزینش و پارامترهای مشترک تحلیل و در نهایت نتایج حاصل با قوانین مربوط به جانمایی چندگانه ترکیب شدند.

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

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

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

3-2-1 مدل گزینش

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

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

استفاده از مدل‌های گزینش در رویارویی با داده‌های 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 توأم 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 می‌کند که این احتمال‌ها در مرحله دوم، تعیین مدل

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

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

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

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 اندازه‌گیری شده برای فرد i و $patches_i$ تعداد بسته‌های نیکوتینی است که شخص در طول مطالعه دریافت کرده است. مدل انصراف استفاده شده نیز مدل رگرسیون لجستیک به صورت زیر بود:

$$\text{Logit}(p_{di}(y_{i,di}, H_{ij})) = \phi_0 + \phi_1 y_{i,di} + \phi_2 y_{i,di-1}$$

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

طولی، به عنوان مخدوش‌گر وارد مدل می‌شوند 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 آن‌ها گسترش مدل حکمن در مطالعات طولی با داده‌های ریزش شده بود که در آن از مدل لجستیک برای مخاطره ریزش استفاده شد. از آن پس مقالات زیادی در اجرا و گسترش مدل‌گزینش دیگل و کنوارد به چاپ رسید. از جمله تروکسل و همکاران این مدل‌ها را برای کنترل گمشدگی‌های غیریکنواخت گسترش دادند، (30). در سال 1995 نیز بیکر، (31) و در همان سال فیتمورس و همکاران، (32)، این مدل‌ها را برای متغیر پاسخ گسسته نیز توسعه دادند. علاوه بر این فعالیت‌هایی نیز در زمینه مطالعات بقا بر اساس مدل‌های گزینش انجام گرفت، (33-35). برای اطلاعات بیشتر در این زمینه، به کتاب فیتموریسو همکاران مراجعه کنید، (6). برای درک بهتر موضوع به کاربرد مدل انتخاب در داده‌های کارآزمایی بالینی ترک سیگار می‌پردازیم.

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

$\phi_1(SD)$	$\phi_1(SD)$	$\beta_3(SD)$	$\beta_2(SD)$	$(SD)\beta_1$
-۰.۰۳	۱.۲۸	-۰.۰۸	۰.۰۲	-۰.۲۸
(۰.۲۳)	(۰.۳۳)	(۰.۰۷)	(۰.۰۵)	(۰.۰۵)

$f(y_i, r_i | X_i, \theta, \phi) = f(y_i | r_i, X_i, \theta) f(r_i | X_i, \phi)$
گسترش مدل‌های گزینش برای کنترل گمشده‌های آگاهی بخش به سال 1977 بر می‌گردد؛ در آن سال رویین روش‌هایی را برای استفاده از پیشین‌های آگاهی بخش در جهت بهره‌برداری از اطلاعات افراد بی‌پاسخ بررسی کرد، (36). لیتل در سال‌های 1993 و 1994 به عنوان راهکار دیگری در مقابل مدل‌های گزینش، گروهی کلی از مدل‌ها را تحت عنوان «مدل‌های الگوی آمیخته» برای تحلیل داده‌های گمشده معرفی و فرمول بندی کرد، (37، 38). در این زمینه قبل از لیتل هم فعالیت‌هایی صورت گرفته بود؛ از جمله توسط مارینی و همکاران، (39)، در سال 1980 و گلین و همکاران، (40)، در سال 1986.

در مدل فوق تنها اثر درمان CM معنی‌دار بود. اثر درمان RP و اثر متقابل بین RP و CM معنی‌دار نشد. ضریب رگرسیونی ϕ_1 نیز به طور معنی‌داری بزرگ‌تر از صفر است ($\phi_1 = 1.28$ و $P = 0.0002$) که بیانگر این مطلب است که هر چه داده گمشده موردنظر بزرگ‌تر باشد، احتمال انصراف بیشتر خواهد بود. به عبارت دیگر انصراف احتمالاً غیرقابل چشم‌پوشی است.

3-2-2- مدل الگوی آمیخته

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

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 می کند، استفاده از آن معمولاً مفیدتر و قابل فهم تر 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 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 مستقل)، مستقیماً در دسترس نیستند. استنباط های 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 x_i استقلال شرطی دارند (به شرط گروهی از پارامترهای

ولی مقاله های لیتل این مدل ها را به طور کامل و به صورت آماری به روشی کلی معرفی کردند. پس از آن مقاله های زیادی در معرفی و گسترش این مدل ها برای داده های طولی به چاپ رسید، (41-43). 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 درستی برای کنترل گمشده های غیرقابل چشم پوشی، 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 آریبی و کارایی برآوردها، حساسیت

مفاهیم گمشدگی و مدل‌های مناسب برای آن‌ها پرداختند.

با توجه به اهمیت موضوع در مقاله حاضر به صورت کاربردی به بیان مفاهیم و مدل‌های آماری موجود در این زمینه پرداخته شد. گمشدگی و پیامدهای ناشی از آن، الگوها و مکانیزم‌های گمشدگی، قابل چشم‌پوشی بودن و مدل‌های مناسب برای آن و در نهایت مدل‌های مناسب برای گمشده‌های غیرقابل چشم‌پوشی موضوع‌های مباحث مطرح شده در مقاله حاضر را در بر می‌گیرند.

در سال‌های اخیر علاوه بر مباحث تئوری، پیشرفت‌های نرم‌افزاری چشم‌گیری نیز در راستای مدل‌سازی داده‌های طولی با حضور گمشدگی، صورت گرفته است. اکثر نرم‌افزارهای آماری از جمله SAS، S-plus، R، SPSS و Stata قابلیت اجرای مدل‌های مربوط به گمشده‌های قابل چشم‌پوشی معرفی شده در این مقاله را دارند، (18). در مقابل در مورد گمشدگی‌های غیرقابل چشم‌پوشی معمولاً نیازمند به کدنویسی‌های پیچیده هستیم. تاکنون تلاش‌هایی در این زمینه در نرم‌افزارهای SAS، R و WinBUGS انجام شده است. (11)

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

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

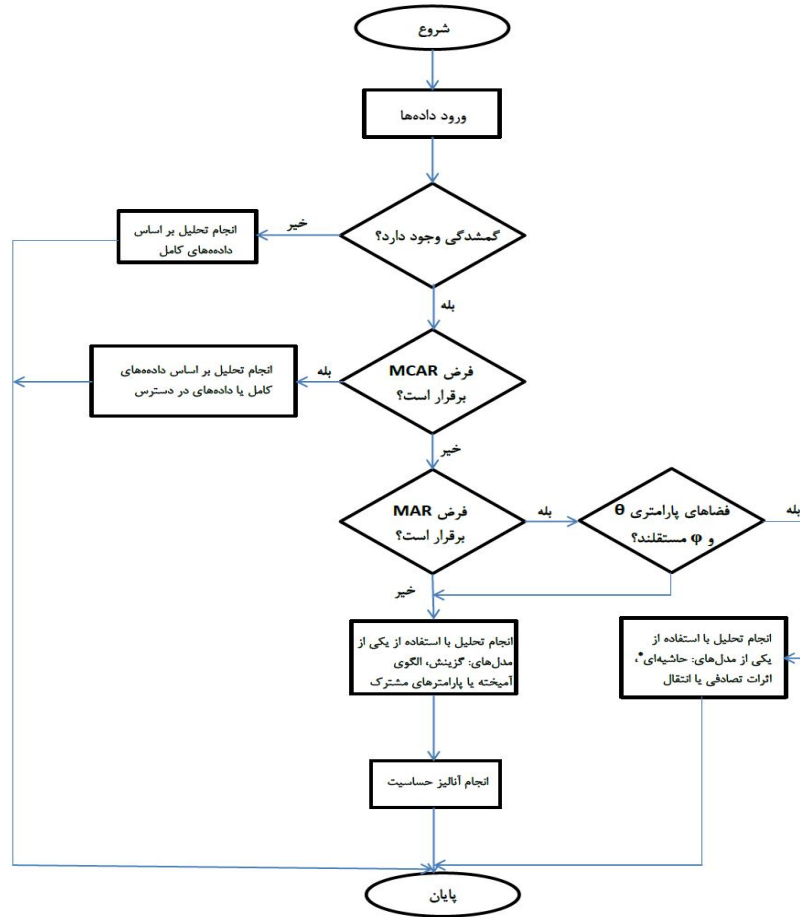
مشترک) و در نتیجه تابع توزیع توأم به صورت تجزیه می‌شود. (16)

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

«پارامترهای» مشترک در این مدل‌ها نقش مخدوش‌کننده را در ارتباط بین r_i و y_i بازی می‌کنند؛ بنا بر این می‌توانند شامل عوامل قابل مشاهده (مانند جنسیت) یا پنهان (یعنی اثرات تصادفی) باشند. این مدل‌ها برای اولین بار توسط وو و کارول، (44) در سال 1988 معرفی شدند. آن‌ها در مطالعه‌ای 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 آریبی در برآوردها می‌شود. بنا بر این چنان‌چه بیان شد، مدل‌های پارامترهای مشترک نیز روشی برای غلبه بر گمشدگی‌های غیرقابل چشم‌پوشی و یا آگاهی بخش خواهند بود. مقاله‌های کلیدی در معرفی و گسترش این مدل‌ها شامل وو و کارول در سال 1988، فلمن و وو، (45)، در سال 1995، وولفسن و همکاران، (46)، در سال 1997 هستند.

بحث و نتیجه‌گیری

مشکل وجود داده‌های گمشده در مطالعات طولی و کارآزمایی‌های بالینی، در سال‌های 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 وریک و مولنبرگز، (47)، 2004 فیتزمریس و همکاران، (18)، 2005 مولنبرگز و وریک، (48)، 2007 مولنبرگز و کنوارد، (49)، 2008 دانیالز و هگان، (11)، و در همان سال فیتزمریس و همکاران، (6)، و در سال 2009 ابراهیم و مولنبرگز، (7)، به طور گسترده به



شکل شماره 1. فلوجارت 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 می‌نمایم.

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, Goetghebuer E. Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. J Appl Statist 2001;50:15-29.
- 4-Robins JM, Rotnitzky A. 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, Shoptaw 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*. 2nd ed. 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. *J Roy 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-likelihood estimation 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-ignorable drop-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

Zayeri F^{*1}, Akbarzadeh Baghban A.R², Kazemzadeh M³, Yaseri M⁴, Abbasi A.M⁵

(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)