مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۱، شماره‌ی ۱، بهار و تابستان ۱۳۸۹، صص ۱۰۹ – ۱۳۶

مقایسه‌ی دو رهیافت برای برخورد با متغیرهای کمکی گم‌شده در رگرسیون لوژستیک

چاو- ییِنگ جُوان پنگ[†] و جین ژو[‡]

[†] دانشگاه ایندیانا، بلومینگتون
[†] شرکت فناوری ژنتیک

مترجم: شیرین گلچی و ساناز مهندسی

پژوهشکده‌ی آمار

چکیده: در طول ۲۵ سال گذشته پیشرفت‌های روش‌شناسانه‌ای در زمینه‌ی تیمار داده‌های گم‌شده صورت گرفته است. بیش‌تر مطالعه‌های انتشاریافته روی داده‌های گم‌شده در متغیرهای وابسته تحت شرایط گوناگون تمرکز داشته‌اند. مطالعه‌ی حاضر در پی آن است که این خلأ را با مقایسه‌ی دو رهیافت برای برخورد با داده‌های گم‌شده در متغیرهای کمکی رسته‌ای در رگرسیون لوژستیک پر کند: روش امید ریاضی- ماکسیمم‌سازی (EM) وزن‌ها و جانهی چندگانه (MI). داده‌های نمونه به‌صورت تصادفی از جامعه‌ای با مشخصه‌های معلوم انتخاب شده‌اند. داده‌های گم‌شده‌ی مربوط به متغیرهای کمکی تحت دو حالت شبیه‌سازی شده‌اند: گم‌شده‌ی کاملاً تصادفی و گم‌شده‌ی تصادفی با نرخ‌های گم‌شدگی متفاوت. یک مدل رگرسیونی لوژستیک با استفاده از یکی از دو رهیافت EM یا MI بر هر نمونه برازش داده شده است. عملکرد این دو رهیافت با چهار معیار مورد مقایسه قرار گرفته است. اریبی، کارایی، پوشش و نرخ عدم پذیرش. نتیجه‌ها به‌طور کلی MI را بر

Peng, C.Y.J. and Zhu, J. (2008). Comparison of Two Approaches for Handling Missing Covariates in Logistic Regression, *Educational and Psychological Measurement*, **68**, 58-77.

دریافت:۱۳۸۸/۲/۲۵ ، پذیرش: ۱۳۸۸/۹/۲

EM ترجیح می‌دادند. مسائل عملی از جمله اجرا، گنجاندن متغیرهای کمکی پیوسته و اثر متقابل بین متغیرهای کمکی مورد بحث قرار گرفته‌اند.

واژگان کلیدی: رگرسیون لوژستیک؛ داده‌های گم‌شده؛ متغیرهای کمکی رسته‌ای گم‌شده؛ روش وزن‌ها؛ جانهی چندگانه؛ الگوریتم EM.

۱- مقدمه

وجود داده‌های گم‌شده در مطالعه‌های روان‌شناختی و آموزشی بیش از آن که استثناء باشد قاعده است. برای مثال، نوشتگان گوناگون نشان می‌دهند که یک نرخ ریزش ۲۰% به‌عنوان معیاری برای مطالعه‌های مربوط به جوانی، برنامه‌های مدرسه- مبنا، و پژوهش‌های بالینی در نظر گرفته می‌شود ([۸]، [۱۱]، [۱۵]). دلیل‌های بسیاری در گم شدن داده‌ها در پروژه‌های تحقیقاتی سهم دارند، از جمله ریزش واحدها در پیش‌آزمون- پس‌آزمون یا طرح‌های طولی، دستورالعمل‌های مبهم یا سئوال‌های مداخله‌گرانه در یک آمارگیری و داده‌دهی ضعیف یا ثبت ناقص آمار ثبتی در داده‌های بایگانی‌شده، که مورد اخیر نشان داده است که اریبی پنهان در داده‌ها وارد می‌کند [۵]. دلیل(های) داده‌های گم‌شده هرچه باشد، اثر آن روی تحقیقات کمّی یک نگرانی بزرگ برای روش‌شناسان بوده است.

در طول ۲۵ سال گذشته پیشرفت‌هایی در زمینه‌ی نظریه‌های مربوط به روش‌های داده‌های گم‌شده و کاربردهای آن در مجموعه‌های داده‌های واقعی صورت گرفته است ([۱۳]، [۲۴]). بیش‌تر مطالعه‌های انتشاریافته بر داده‌های گم‌شده در متغیرهای وابسته تحت حالت‌های مختلف تمرکز داشته‌اند ([۱۲]، [۲۱])، هرچند روش‌های تیمار مقادیر گم‌شده در متغیرهای مستقل (یا متغیرهای کمکی) تفاوت اساسی با روش‌های تیمار مقادیر گم‌شده در متغیرهای وابسته ندارند. تحت فرض‌های آماری حتمی، می‌توان توسط روش ساده‌ای چون حذف موردی به شکلی کارامد با داده‌های گم‌شده در یک متغیر مستقل برخورد کرد [۷].

ابراهیم [۱۰] یک روش امید ریاضی- ماکسیمم‌سازی (EM) وزن‌ها را برای به دست آوردن براوردهای ماکسیمم درستنمایی (ML) ضریب‌های رگرسیونی برای مدل‌های خطی تعمیم‌یافته با متغیرهای کمکی رسته‌ای پیشنهاد کرد. هورتون و لرد [۹] روش ابراهیم را داخل یک الگوریتم اجرا کردند و آن را در چندین مثال سلامت عمومی نشان دادند.

مقاله‌ی آن‌ها بسط‌ها و محدودیت‌های روش ابراهیم را بدون مقابله با یک روش دیگر مانند جانهی چندگانه (MI) مورد بحث قرار می‌داد. لیتل [۱۲] روش‌های متفاوت برخورد با مقادیر گم‌شده در متغیرهای کمکی را در تحلیل رگرسیونی مورد بررسی قرار داد و نتیجه گرفت که روش‌های ارجح روش‌های براورد مدل- مبنا هستند (یعنی، روش‌های ML). شِیفر و گراهام ([۲۳]، ص ۱۷۳) هر دوی روش‌های مبتنی بر درستنمایی و روش‌های پارامتری MI را به‌عنوان نمایانگر آخرین تحولات عملی تحت فرض «گم‌شده‌ی تصادفی» در نظر می‌گیرند. در همان مقاله شِیفر و گراهام پیشنهاد می‌کنند که پژوهش‌های بیش‌تری برای مقایسه‌ی رسمی عملکرد MI با روش‌های وزن‌دهی مانند روش EM ابراهیم انجام شود. دیدلز [۴] براوردهای ML ضریب‌های رگرسیونی لوژستیک با مقادیر گم‌شده در متغیرهای کمکی را هنگامی که توزیع این متغیرها نامشخص است مورد بررسی قرار داد. کار او چنین نتیجه‌گیری می‌کرد که این رهیافت پارامتری در صورتی که توزیع مفروض شدیداً از توزیع واقعی انحراف داشته باشد می‌تواند به اریبی جدی منجر شود.

راگوناتان [۱۷] بزرگی اریبی حاصل از استفاده از (الف) یک روش حذف فهرست‌وار، (ب) یک فن وزن‌دهی و (پ) یک MI برای برخورد با داده‌های ناقص در تحلیل رگرسیونی لوژستیک را مورد مقایسه قرار داد. نتیجه‌های شبیه‌سازی مبتنی بر ۲۵۰۰ نمونه به اندازه‌ی ۱۰۰۰ نشان می‌داد که همان‌گونه که انتظار می‌رفت اریبی روش حذف فهرست‌وار از همه بیش‌تر بود. روش وزنی و MI هر دو براوردهایی تولید کردند که توزیع‌های نمونه‌گیری آن‌ها روی مقدار واقعی جامعه، ۰/۵، متمرکز بود. راگوناتان با توجه به این‌که (الف) روش وزنی رهیافت ساده‌ای برای تصحیح اریبی بود، اما به اندازه‌ی MI کارا نبود؛ (ب) MI کاربردی‌تر و کاراتر از روش وزنی بود؛ و (پ) هر دو روش تنها تحت فرض گم‌شده‌ی تصادفی معتبر بودند، روی این نتیجه‌ها اظهار نظر کرد. او روش ML را به‌عنوان رهیافتی دیگر برای مجموعه‌ی محدودی از مدل‌ها، مانند مدل‌های رگرسیونی لوژستیک، مورد بحث قرار داد و به غیر عملی بودن به کارگیری روش ML به خاطر دشواری‌های فنی آن اشاره کرد. متأسفانه او در عمل روش ML را برای داده‌های شبیه‌سازی‌شده به کار نگرفت؛ لذا در مطالعه‌ی وی هیچ مقایسه‌ای بین ML و MI صورت نگرفته است.

مطالعه‌ی حاضر در پی آن است که خلأ موجود در نوشتگان را با مقایسه‌ی مستقیم و رسمی رهیافت MI با روش وزنی EM (یک رهیافت ML) پر کند. به‌طور خاص این دو

رهیافت برای برخورد با داده‌های گم‌شده در متغیرهای کمکی رسته‌ای در رگرسیون لوژستیک مورد مقایسه قرار گرفته‌اند. در ابتدا داده‌های تصادفی در جامعه‌ای با مشخصه‌های معلوم انتخاب شدند. سپس داده‌های گم‌شده در متغیرهای کمکی تحت دو حالت: گم‌شده‌ی کاملاً تصادفی (MCAR) و گم‌شدگی تصادفی (MAR) با نرخ‌های گم‌شدگی متفاوت شبیه‌سازی شدند. یک مدل رگرسیونی لوژستیک با استفاده از یکی از دو رهیافت EM یا MI برای برخورد با داده‌های گم‌شده در متغیرهای کمکی بر هر نمونه برازش داده شده است. براورد پارامترها و فاصله‌های اطمینان آن‌ها تحت هر رهیافت محاسبه شده است. عملکرد این دو رهیافت روی چهار معیار: اریبی، کارایی، پوشش و نرخ عدم پذیرش ارزشیابی شده است. چون تحلیل موردی کامل (CC) تحت حالت MCAR معتبر است ([۱۲] و [۱۳])، نتیجه‌های آن ارایه شده و به‌عنوان مرجعی مورد استفاده قرار گرفته است که این دو رهیافت نسبت به آن مقایسه شده‌اند. یافته‌های حاصل از این مطالعه‌ی شبیه‌سازی می‌تواند کار هورتون و لرد [۹] را در جهت استفاده از رهیافت MI علاوه بر روش وزنی EM برای تیمار متغیرهای کمکی رسته‌ای گم‌شده در رگرسیون لوژستیک بسط دهد. یافته‌ها نتیجه‌های گراهام و شِیفر [۷] را نیز با بررسی عملکرد MI برای هر دو مجموعه‌ی داده‌های بزرگ و کوچک با مقادیر گم‌شده در متغیرهای کمکی رسته‌ای، مورد تمجید قرار می‌دهند.

۲- اصطلاح‌شناسی و روش

در این بخش دو عبارت را که به فراوانی در نوشتگان داده‌های گم‌شده مورد استفاده قرار می‌گیرند تعریف می‌کنیم: MCAR و MAR. این‌ها به مکانیزمی که به داده‌های گم‌شده منجر شده است اشاره دارند [۱۳]. دو رهیافت داده‌های گم‌شده (روش وزنی EM و MI) نیز به‌همین ترتیب تعریف شده و مورد بررسی قرار گرفته‌اند. این چهار عبارت برای درک مطالعه‌ی حاضر از اهمیت اساسی برخوردارند.

MCAR

همان‌طور که توسط لیتل و روبین [۱۳] تعریف شده است، «اگر احتمال یک پاسخ نه به مقدار مشاهده‌شده بستگی داشته باشد و نه به مقدار گم‌شده‌ای که می‌توانست گرداوری یا ثبت شود، داده‌های گم‌شده داده‌های گم‌شده‌ی کاملاً تصادفی هستند» (ص ۱۴). به‌نحوی

رسمی‌تر فرض کنید Y نمایانگر یک ماتریس داده‌ی $n \times p$ تشکیل شده از دو بخش باشد: داده‌هایی که کاملاً مشاهده شده‌اند و داده‌هایی که به‌صورت بالقوه گم شده‌اند. به عبارت دیگر $\boldsymbol{Y} = (\boldsymbol{Y}_{\text{observed}}, \boldsymbol{Y}_{\text{missing}})$. فرض کنید $\boldsymbol{R}$ یک ماتریس $n \times p$ از مقادیر صفر و ۱ متناظر با $\boldsymbol{Y}$ مشاهده‌شده یا گم‌شده باشد. روبین [۱۸] MCAR را به‌صورت زیر تعریف کرده است:

$$(۱) \qquad \text{Probability}(\boldsymbol{R} \mid \boldsymbol{Y}_{\text{missing}}, \boldsymbol{Y}_{\text{observed}}, \zeta) = \text{Probability}(\boldsymbol{R} \mid \zeta)$$

به عبارت دیگر اگر احتمال این‌که $\boldsymbol{Y}$ گم‌شده باشد به خود مقدار گم‌شده، یا مقادیر مشاهده‌شده‌ی $\boldsymbol{Y}$ بستگی نداشته، بلکه در عوض به برخی پارامترهای نامعلوم ζ بستگی داشته باشد، آن‌گاه گفته می‌شود که داده‌ها MCAR هستند. تحت حالت MCAR، داده‌های گم‌شده را می‌توان به‌شکل یک زیرنمونه‌ی تصادفی از داده‌های بالقوه‌ی کامل، تیمار کرد؛ لذا رهیافت CC یا حذف فهرست‌وار، در این حالت معتبر است. به بیان رسمی، مکانیزم گم‌شدن داده‌ها که دلیل‌های گم شدن داده‌ها را در بر می‌گیرد می‌تواند در استنباط مبتنی بر نمونه‌گیری و درستنمایی تحت MCAR نادیده گرفته شود ([۱۳]، ص۱۵).

MAR

فرض MAR اظهار می‌کند که:

$$(۲) \qquad \text{Probability}(\boldsymbol{R} \mid \boldsymbol{Y}_{\text{missing}}, \boldsymbol{Y}_{\text{observed}}, \zeta) = \text{Probability}(\boldsymbol{R} \mid \boldsymbol{Y}_{\text{observed}}, \zeta)$$

معادله‌ی (۲) حاکی از آن است که احتمال شرطی گم‌شده بودن $\boldsymbol{Y}$ به شرط هر دوی $\boldsymbol{Y}_{\text{missing}}$ و $\boldsymbol{Y}_{\text{observed}}$ برابر است با احتمال شرطی مقادیر گم‌شده‌ی $\boldsymbol{Y}$ به شرط مقادیر مشاهده‌شده‌ی $\boldsymbol{Y}$. به بیان دیگر، MAR به این معنی است که احتمال گم‌شدن یک مشاهده ممکن است به مقادیر مشاهده‌شده بستگی داشته اما به خود مقدار گم‌شده بستگی نداشته باشد.

محدودیت MAR از MCAR کم‌تر است؛ لذا از MCAR به‌عنوان حالت خاص MAR یاد می‌شود. تحت حالت MAR، مکانیزم گم‌شدن برای استنباط مبتنی بر درستنمایی قابل چشم‌پوشی است ([۱۳]، ص ۱۵). اساساً قابل چشم‌پوشی به این معنی

است که لازم نیست مکانیزم گم‌شدن داده‌ها به‌صورت بخشی از فرایند براورد مدل‌سازی شود. دو رهیافت مورد بررسی در این مقاله تحت هر دو حالت MCAR یا MAR کاربرد دارند و در نوشتگان نشان داده شده است که هر دوی این رهیافت‌ها تحت حالت گم شدن غیر تصادفی بهتر از رهیافت CC یعنی حذف فهرست‌وار، عمل می‌کنند.

روش وزنی EM

ابراهیم [۱۰] یک روش براورد ML برای مدل‌های رگرسیونی خطی تعمیم‌یافته با متغیرهای کمکی رسته‌ای گم‌شده پیشنهاد کرده است. این کاربردی از الگوریتم EM همراه با روش وزنی است. الگوریتم EM یک رهیافت اصولی است که دو مرحله را به‌صورت متوالی تکرار می‌کند: مرحله‌ی E و مرحله‌ی M ([۳] و [۱۳]). در مرحله‌ی E، امید ریاضی لگاریتم درستنمایی داده‌های کامل نسبت به توزیع شرطی داده‌های گم‌شده به شرط داده‌های مشاهده‌شده و براورد پارامترهای به دست آمده از تکرار قبل گرفته می‌شود. در مرحله‌ی M، براوردهای جدید پارامتر توسط ماکسیمم کردن لگاریتم درستنمایی مرحله‌ی E به دست می‌آیند، و براوردهای جدید پارامتر در تکرار بعدی مرحله‌ی E مورد استفاده قرار می‌گیرند. الگوریتم EM تا زمانی که درستنمایی‌های مشاهده‌شده‌ی تولید شده در دو تکرار متوالی تقریباً معادل باشند ادامه پیدا می‌کند. در تکرار آخر که الگوریتم همگرا می‌شود، پارامترهای براورد شده نهایی می‌شوند.

الگوریتم EM می‌تواند با وزن‌دهی مکرر مشاهده‌ها، مانند روش ابراهیم [۱۰]، اجرا شود. به خصوص وزن‌ها در مرحله‌ی E به‌منظور محاسبه‌ی لگاریتم درستنمایی با مقادیر گم‌شده در متغیرهای کمکی داخل یک مجموعه‌ی داده‌های ناقص قرار داده می‌شوند. فرض کنید یک مجموعه‌ی داده‌های ناقص دارای یک متغیر وابسته‌ی دو حالتی (Y) و سه متغیر کمکی دو حالتی (X_1, X_2, X_3) است، که در سمت چپ جدول ۱ ارایه شده‌اند. این مجموعه‌ی داده‌های ناقص دارای دو الگوی داده‌های گم‌شده است: یکی فاقد مقدار X_1 و دیگری فاقد مقادیر X_1 تا X_3 است. برای این دو الگوی داده‌های گم‌شده، وزن‌هایی برای افزوده شدن به مجموعه‌ی داده‌ها درج می‌شوند (سمت راست جدول ۱ را ببینید). اگر مشاهده‌ای هیچ مقدار گم‌شده‌ای نداشته باشد، وزن منتسب‌شده برابر ۱ است، مانند حالت‌های ۱، ۲، ۴، ۵ و ۶. در صورتی که مشاهده‌ای دارای یک مقدار گم‌شده باشد (مانند حالت‌های ۳ و ۷) وزن به‌صورت یک احتمال شرطی براورد می‌شود. بنا بر این،

جدول ۱- تصویری از روش امید ریاضی- ماکسیمم‌سازی وزن‌ها توسط هورتون و لِرد

مجموعه‌ی داده‌های افزوده						مجموعه‌ی داده‌های اصلی[الف]				
wt	X_3	X_2	X_1	Y	حالت	X_3	X_2	X_1	Y	حالت
۱	۰	۰	۰	۰	۱	۰	۰	۰	۰	۱
۱	۱	۰	۰	۰	۲	۱	۰	۰	۰	۲
w۳۱	۰	۰	۰	۰	۳	۰	—	۰	۰	۳
w۳۲	۰	۰	۱	۰	۳					
1	۰	۱	۰	۰	۴	۰	۱	۰	۰	۴
۱	۱	۰	۱	۱	۵	۱	۰	۱	۱	۵
۱	۰	۰	۰	۱	۶	۰	۰	۰	۱	۶
w۷۱	۰	۰	۰	۱	۷	—	—	—	۱	۷
w۷۲	۱	۰	۰	۱	۷					
w۷۳	۰	۱	۰	۱	۷					
w۷۴	۱	۱	۰	۱	۷					
w۷۵	۰	۰	۱	۱	۷					
w۷۶	۱	۰	۱	۱	۷					
w۷۷	۰	۱	۱	۱	۷					
w۷۸	۱	۱	۱	۱	۷					

الف. داده‌های گم‌شده با خط تیره نشان داده شده‌اند.

$$w\,۳۱ = P(x_۱ = ۰ \mid x_۲ = x_۳ = y\ = ۰, \hat{\theta})$$

و

$$w\,۳۲ = P(x_۱ = ۱ \mid x_۲ = x_۳ = y\ = ۰, \hat{\theta})$$

که در آن $\hat{\theta}$ براورد پارامتر به دست آمده از مرحله‌ی M قبلی است. وزن‌های منتسب به یک مشاهده باید تابع این قید باشند که جمع آن‌ها برابر ۱ شود. لذا $w\,۳۱ + w\,۳۲ = ۱$ و $\sum_{i=۱}^{۸} w\,۷i = ۱$. لگاریتم درستنمایی مجموعه‌ی داده‌های افزوده سپس در مرحله‌ی بعدی E تعیین می‌شود. الگوریتم EM به کار گرفته شده روی مجموعه‌ی داده‌های افزوده، بین مرحله‌های E و M تکرار می‌شود تا زمانی که براوردهای پارامترهای مورد نظر نهایی شوند.

روش EM وزن‌های ابراهیم به‌طور گسترده توسط هورتون و لِرد بررسی و نشان داده شده است [۹]. همان‌طور که توسط هورتون و لِرد (۱۹۹۸، ص ۴۰) نشان داده شده است. خطاهای استاندارد نمونه‌ی بزرگ براوردهای پارامتر می‌توانند با استفاده از روش لوییس [۱۴] محاسبه شوند. اجرای روش وزنی EM ابراهیم توسط هورتون و لِرد برای مطالعه‌ی حاضر تصحیح و تطبیق داده شده است.

MI

MI یک فن مبتنی بر شبیه‌سازی است که مقادیر گم‌شده را چندین بار جانهی می‌کند [۱۹]. برخلاف روش جانهی تکی، MI یک معیار عدم قطعیت فراهم می‌کند که به‌وسیله‌ی آن داده‌های گم‌شده از روی داده‌های مشاهده‌شده پیش‌بینی می‌شوند. به خاطر سادگی و کلیت آن، MI به‌صورت یک روش محبوب در میان محققان علوم اجتماعی در آمده است. برای مقدمه‌ای بر MI مقاله‌ی [۲۲] را توصیه می‌کنیم.

ایده‌ی اولیه‌ی MI پر کردن مقادیر گم‌شده به دفعات متعدد به‌منظور ساختن مجموعه‌های داده‌های کامل چندگانه است. از لحاظ تحلیلی، MI شامل سه مرحله است: جانهی، تحلیل و ادغام. در مرحله‌ی اول، جانهی، هر مقدار گم‌شده توسط $m > ۱$ مقدار شبیه‌سازی‌شده جایگزین می‌شود. در فرایند پر کردن، یک توزیع توأم از داده‌های کامل (مشاهده‌شده و گم‌شده) و یک توزیع پیشین از پارامترها برای الگوریتم داده‌افزایی به‌منظور شبیه‌سازی عددهای تصادفی برای داده‌های گم‌شده، فرض می‌شوند. برای مجموعه‌های داده‌های دارای متغیرهای رسته‌ای، یک توزیع چندجمله‌ای برای داده‌های کامل و یک توزیع پیشین دیریکله برای پارامترها فرض می‌شود. تحت فرض قابل چشم‌پوشی MAR، می‌توان m عدد تصادفی مستقل از توزیع شرطی ایستای داده‌های گم‌شده، به شرط داده‌های مشاهده‌شده، مطابق رهیافت بیزی، شبیه‌سازی کرد. بعد از مرحله‌ی جانهی، m مجموعه‌ی داده‌های کامل ساخته می‌شود. در مرحله‌ی دوم، تحلیل، هر یک از m مجموعه‌ی داده‌های کامل توسط یک روش داده‌های کامل استاندارد، مانند تحلیل رگرسیونی معمولی کم‌ترین توان‌های دوم، تحلیل می‌شود. سرانجام در مرحله‌ی سوم، ادغام، نتایج m تحلیل با استفاده از قاعده‌ی رابین (۱۹۸۷، نشان داده‌شده در [۱۶]) ادغام می‌شوند تا یک نتیجه‌ی نهایی، مثلاً یک براورد فاصله‌ای از ضرایب رگرسیون، یک p– مقدار از آزمون فرض صفر، یا یک آماره‌ی آزمون نسبت درستنمایی به دست آید.

تعداد مجموعه‌های داده‌های جانهی‌شده لازم نیست خیلی زیاد باشد. معمولاً ۳ تا ۱۰ مجموعه‌ی داده‌های جانهی‌شده نتایج مطلوبی ارایه می‌دهند ([۱۶]، [۲۱]، [۲۲] و [۲۴]).

طرح

هدف اولیه‌ی این مطالعه مقایسه‌ی روش وزنی ابراهیم (۱۹۹۰) (یک رهیافت EM) با رهیافت MI برای برخورد با داده‌های گم‌شده در متغیرهای کمکی رسته‌ای مدل‌های رگرسیونی لوژستیک بوده است. برای مطالعه‌ی عملکرد EM و MI پنج مرحله طراحی و اجرا شده است. در مرحله‌ی اول، یک جامعه‌ی تجربی تعریف شده است. یک مدل رگرسیون لوژستیک به داده‌های جامعه برازش داده شده و لذا ضرایب رگرسیون لوژستیک حقیقی تولید شده‌اند. در مرحله‌ی دوم، ۱۰۰۰ نمونه‌ی تصادفی به دو اندازه از جامعه انتخاب شده‌اند. در مرحله‌ی سوم، داده‌های گم‌شده در نمونه تحت شرایط MCAR و MAR با نرخ‌های گم‌شدگی مختلف به وجود آمده‌اند. در مرحله‌ی چهارم، همان مدل رگرسیون لوژستیک برازش داده‌شده به جامعه برای هر نمونه به کار رفته است؛ ضریب‌های رگرسیونی و خطاهای استاندارد آن‌ها با استفاده از یکی از دو رهیافت EM یا MI برای تیمار داده‌های گم‌شده در نمونه، براورد شده‌اند و در مرحله‌ی پنجم، نتیجه‌ها برای ۱۰۰۰ نمونه‌ی شبیه‌سازی‌شده تحت هر حالت گرداوری شده است. عملکرد EM و MI روی چهار معیار اریبی، کارایی، پوشش و نرخ‌های عدم پذیرش ارزیابی شده است. به‌علاوه، نتیجه‌های تحلیل CC ارایه شده و به‌عنوان مرجع‌هایی به کار رفته‌اند که این دو رهیافت نسبت به آن‌ها مقایسه شده‌اند.

جامعه

مجموعه داده‌هایی که شامل اطلاعاتی پیرامون رضایت مشتریان از خدمات یک کارت اعتباری به خصوص می‌شد به‌عنوان جامعه به کار رفته است. داده‌ها شامل ۲۱۵۰۴ مشتری (یا صاحب کارت اعتباری) با اطلاعات کامل در مورد ۲۰ متغیر است. دو نوع متغیر جمع‌آوری شده است: (a) پاسخ‌های مشتریان به سئوال‌های آمارگیری (b) متغیرهای متعلق به وجهه‌ی اعتباری شخصی. سئوال‌های آمارگیری اطلاعاتی در مورد استفاده‌ی مشتری از کارت اعتباری می‌پرسیدند، مانند «در سال گذشته چند بار برای

کارت اعتباری درخواست کردید؟». نوع دوم متغیرها در جستجوی اطلاعاتی درباره‌ی وجهه‌ی اعتباری شخصی بودند، از قبیل محدودیت اولیه‌ی اعتبار. ما چهار متغیر کمکی $(X_1 - X_4)$ را که ممکن است روی رضایت مشتریان از خدمات کارت اعتباری، یعنی متغیر وابسته‌ی Y، تأثیر بگذارد انتخاب کردیم. برای متغیرهای Y، X_1 تا X_4، پاسخ مورد نظر با مقدار ۱ و پاسخ دیگر با صفر کدبندی شده است. به‌علاوه یک متغیر کمکی پنجم، X_5، کاملاً نامرتبط با Y، X_1 تا X_4 به‌صورت تصادفی از توزیع برنولی با احتمال (P)، ۰/۲ یا ۰/۴ تولید شده است. این متغیر مستقل پنجم برای ارزیابی عملکرد الگوریتم‌های EM و MI و نسبت به نرخ خطای نوع I مورد استفاده قرار گرفته است. انتخاب این دو پارامتر برنولی توجیه نظری ندارد، اگرچه ۰/۲ افراطی تر از ۰/۴ است.

برای به دست آوردن پارامترهای واقعی جامعه، یک مدل رگرسیون لوژستیک با ۵ متغیر کمکی به داده‌های جامعه برازش داده شد. همه‌ی ضریب‌های رگرسیونی در سطح $\alpha = ۰/۵$ یا کوچک‌تر از نظر آماری معنی‌دار بودند، به جز β_5 که امری قابل درک است. ضریب‌های همبستگی پی‌یرسون بین متغیر وابسته و ۵ متغیر کمکی نشان داد که رضایت مشتریان از یک کارت اعتباری به خصوص (Y) دارای همبستگی مثبت با حد اولیه‌ی اعتبار (X_3)، و همبستگی منفی با تعداد دفعاتی که در سال گذشته برای کارت اعتباری درخواست داده‌اند (X_1) و تعداد سال‌های داشتن کارت اعتباری (X_2) و مجموع حد اعتبار همه‌ی کارت‌های اعتباری آن‌ها (X_4) است. رضایت مشتریان به متغیر کمکی تصادفی برنولی ارتباط نداشت، که نباید داشته باشد. نتیجه‌های رگرسیون لوژستیک و ضریب‌های همبستگی پی‌یرسون را می‌توان از مؤلف اول به‌دست آورد.

۳- روش نمونه‌گیری

به‌محض این که جامعه مشخص شد، ۱۰۰۰ نمونه‌ی تصادفی به اندازه‌ی ۲۰۰ یا $n = ۴۰۰$ انتخاب شدند. نرخ‌های گم‌شدگی و مکانیزم‌ها برای هر نمونه‌ی تصادفی دست‌کاری شدند. از آن‌جا که بیش‌تر ارزیابی‌های MI بر اساس نمونه‌های بزرگ بوده (به‌عنوان مثال، [۶]) و نظریه‌ی آماری حول MI و روش وزنی EM تا اندازه‌ای بر اساس تقریب‌های نمونه‌ی بزرگ است، دو اندازه‌ی نمونه تعیین شدند تا به پیروی از طرح مطالعه‌ی [۷] نمایانگر نمونه‌های بزرگ و کوچک باشند. اندازه‌های نمونه‌ی کوچک‌تر از

۲۰۰ با توجه به نرخ‌های گم‌شدگی و مکانیزم‌های گم‌شدن انجام‌شده در این مطالعه میسر نبودند.

نرخ‌های گم‌شدگی و مکانیزم‌های گم شدن

مقادیر گم‌شده در $X_۳$ طبق یکی از دو مکانیزم MCAR یا MAR تولید شدند. ۴ متغیر کمکی دیگر دست نخورده نگاه داشته شدند. برای مکانیزم MCAR متغیر کمکی $X_۳$ دارای نرخ گم‌شدگی ۴۰% در هر نمونه بود. این حالت از این پس MCAR(۰/۴)خوانده می‌شود. برای مکانیزم MAR دو نرخ گم‌شدگی ایجاد شد. اولی ۱۰% داده‌ی گم‌شده در $X_۳$ تولید می‌کرد اگر متغیر وابسته‌ی Y کد ۱ گرفته بود و ۲۰% اگر Y کد صفر گرفته بود. این الگوی داده‌های گم‌شده از این پس MAR(۰/۱،۰/۲) خوانده می‌شود. نرخ گم‌شدگی دوم در $X_۳$ بیش‌تر بود، یعنی ۳۰% و ۴۰% به ترتیب به ازای Y با کد ۱ یا صفر. این حالت MAR(۰/۳،۰/۴) خوانده می‌شود. ترکیب دو اندازه‌ی نمونه، سه الگوی گم‌شدگی، و دو احتمال برای متغیر تصادفی برنولی ($X_۵$) ، ۱۲ حالت به دست می‌دهد. هر حالت ۱۰۰۰ بار شبیه‌سازی شده است. هر نمونه‌ی شبیه‌سازی‌شده توسط هر دو روش EM و MI تحلیل شده‌اند تا براورد پارامترها و خطاهای استاندارد آن‌ها به دست آید.

۴- تحلیل و جانهی

رهیافت EM با استفاده از برنامه‌ی S-Plus نوشته شده توسط هورتون و لِرد [۹] برای هر نمونه‌ی شبیه‌سازی‌شده به کار رفته است. این برنامه روش وزن‌های EM را برای به دست آوردن براوردهای ML ضریب‌های رگرسیون لوژستیک با ۵ متغیر کمکی $X_۱$ تا $X_۵$ اجرا می‌کند. ضریب‌های رگرسیونی براورد شده و خطاهای استاندارد آن‌ها به دست آمده و بین ۱۰۰۰ نمونه‌ی شبیه‌سازی‌شده همگردانی شده است.

رهیافت MI با استفاده از برنامه‌ی CAT شِیفر، که درS-Plus نیز نوشته شده (قابل دسترسی از نشانی http://www.stat.psu.edu/~jls/misoftwa.html) به کار گرفته شده است. پیش از مدل‌بندی رگرسیون لوژستیک، مقادیر گم‌شده در هر نمونه $m = ۱۰$ بار جانهی شده‌اند. جانهی بر اساس یک مدل چندجمله‌ای برای $X_۱$ تا $X_۵$ و Y بوده است. ده مجموعه از ضریب‌ها براورد شده و خطاهای استاندارد آن‌ها با توجه

به قاعده‌ی رابین [۱۹] ترکیب شده‌اند. انتخاب $m = ۱۰$ مناسب به نظر می‌رسد زیرا کارایی براورد پارامترها بر اساس m جانهی عبارت است از $\left(۱+\frac{\gamma}{m}\right)^{-۱}$، که در آن γ نرخ اطلاعات گم‌شده است ([۱۹]، ص ۱۱۴) برای نرخ‌های پایین داده‌های گم‌شده (۲۰% یا کم‌تر)، بیش‌تر از $m = ۳$ جانهی لازم نیست تا حد اقل ۹۴% کارا باشد.

نتیجه‌های مدل‌بندی رگرسیون لوژستیک با استفاده از هر دو رهیافت EM و MI برای ۱۰۰۰ نمونه‌ی شبیه‌سازی‌شده تحت هر یک از ۱۲ حالت ثبت شده است. نتیجه‌ها شامل براوردهای ضریب‌های رگرسیونی، فاصله‌های اطمینان ۹۵% براوردها و آزمون t براوردها در سطح $\alpha = ۰/۰۵$ هستند.

۵- معیار عملکرد

عملکرد رهیافت‌های EM و MI از طریق ۴ معیار ارزیابی شد: اریبی، کارایی، پوشش، و نرخ عدم پذیرش. این چهار معیار [۷] برای ارزیابی عملکرد MI در مجموعه‌های کوچک داده‌های چندمتغیره و [۲۳] برای بحث عملکرد نسبی روش‌های گوناگون داده‌های گم‌شده شامل حذف فهرست‌وار، جانهی‌های تکی، ML و MI به کار رفته‌اند. این چهار معیار به‌صورت عملیاتی برای این مطالعه به‌صورت زیر تعریف شده‌اند:

اریبی: اریبی به‌صورت تفاضل بین میانگین ۱۰۰۰ براورد ضریب و ضریب جامعه‌ی متناظر تعریف شده است. یک رهیافت بهتر مقدار جامعه را به‌طور متوسط با اریبی کم‌تری تولید می‌کند.

کارایی: کارایی به‌صورت تغییرپذیری براوردها حول ضریب واقعی جامعه تعریف شده است. کارایی در این مطالعه توسط پهنای متوسط فاصله‌ی اطمینان ۹۵% اندازه‌گیری می‌شود. فاصله‌ی ۹۵% تقریباً ۴ برابر مقدار خطای استاندارد است. لذا، فاصله‌ی پهن‌تر حاکی از رهیافتی با کارایی کم‌تر است.

پوشش: پوشش به‌صورت درصد فاصله‌های اطمینان ۹۵% است که شامل ضریب واقعی جامعه در بین ۱۰۰۰ فاصله‌ی اطمینان این چنین است. یک رهیافت بهتر

باید فاصله‌های اطمینان ۹۵ درصدی تولید کند که در حدود ۹۵% اوقات شامل پارامتر مزبور باشند.

نرخ عدم پذیرش: نرخ عدم پذیرش به‌صورت درصد دفعاتی که فرض صفر درباره‌ی ضریب رگرسیونی جامعه‌ی X_5 در میان ۱۰۰۰ نمونه رد شده است، تعریف می‌شود. چون X_5 عمداً به‌گونه‌ای تولید شده است که با چهار متغیر کمکی دیگر و Y ناهمبسته باشد، ضریب رگرسیونی واقعی جامعه برای X_5 در واقع صفر است. لذا، فرض صفر ضریب رگرسیونی جامعه‌ی X_5 برابر صفر باید ۵ درصد اوقات در سطح α برابر ۰/۰۵ رد شود.

زمانی که اندازه‌ی نمونه از ۲۰۰ به ۴۰۰ افزایش می‌یافت انتظار می‌رفت که کارایی EM و MI طبق تعریف آماری استاندارد «کارایی»، افزایش یابد و آزمون t مرتبط با این براوردها پرتوان‌تر گردد.

۶- نتیجه‌ها

عملکرد EM و MI از لحاظ اریبی، کارایی، پوشش و نرخ عدم پذیرش در جدول‌های ۲ و ۳ برای پارامتر دو جمله‌ای $p = ۰/۲$ ارایه شده است. به خاطر تشابه یافته‌ها، نتیجه‌ها برای پارامتر دوجمله‌ای $p = ۰/۴$ ارایه نشده‌اند اما از مؤلف اول در دسترس‌اند. تفاوت نتیجه‌ها بین $p = ۰/۲$ و $p = ۰/۴$ در بخش ۷ (بحث) ذکر شده است. جدول ۲ نتیجه‌ها را برای $n = ۴۰۰$ و سه الگوی داده‌های گم‌شده: MAR(۰/۱,۰/۲)، MAR(۰/۳,۰/۴) و MCAR(۰/۴) به‌ترتیب نشان می‌دهد. جدول ۳ نتیجه‌ها را برای $n = ۲۰۰$ و همان سه مکانیزم داده‌های گم‌شده نشان می‌دهد. دو متغیر X_3 و X_5 در هر جدول مشخص شده‌اند زیرا X_3 شامل داده‌های گم‌شده و X_5 متغیر تصادفی دوجمله‌ای است.

مقایسه‌ی EM و MI در مقابل CC

اریبی: برای هر متغیر کمکی و عرض از مبدأ در جدول‌های ۲ و۳، بزرگ‌ترین اریبی که بدترین اریبی هم هست مشخص شده است. برای هر سه رهیافت، زمانی که نرخ گم‌شدگی بالا است، در MAR (۰/۳,۰/۴) ، نسبت به زمانی که نرخ گم‌شدگی پایین است، در MAR (۰/۱,۰/۲) ، اریبی بدتر است. رهیافت CC بدترین اریبی را با بیش‌ترین

جدول ۲- اریبی، کارایی، پوشش و نرخ عدم پذیرش روش‌های EM، MI، و CC برای $n = ۴۰۰$ و p برنولی = ۰/۲

		اریبی			کارایی			پوشش سطح واقعی = ۹۵٪			نرخ عدم پذیرش [الف] مقدار اسمی $\alpha = ۰/۰۵$		
		EM	MI	CC	EM	MI	CC	EM	MI	CC	EM	MI	CC
MAR(۰/۱،۰/۲)	عرض از مبدا	-۰/۰۳۱۱	۰/۰۳۹۹	۰/۰۵۴۲	۱/۳۰۰	۱/۲۱۲	۱/۲۸۶	۰/۹۷۰	۰/۹۶۶	۰/۹۴۴			
	X_1	-۰/۰۰۸۱	-۰/۰۱۴۷	۰/۰۱۵۲	۱/۲۷۰	۱/۱۹۰	۱/۲۷۵	۰/۹۶۵	۰/۹۵۱	۰/۹۵۳			
	X_2	-۰/۰۱۰۶	۰/۰۰۰۳	-۰/۰۴۷۴	۱/۴۴۵	۱/۳۳۴	۱/۴۴۴	۰/۹۵۸	۰/۹۴۲	۰/۹۶۲			
	X_3	۰/۰۱۳۸	-۰/۱۰۰۵	۰/۰۳۳۷	۱/۴۴۶	۱/۳۲۲	۱/۲۹۸	۰/۹۷۱	۰/۹۷۸	۰/۹۵۲			
	X_4	-۰/۰۰۹۰	۰/۰۰۷۰	-۰/۰۰۲۳	۱/۲۵۰	۱/۱۶۹	۱/۲۵۶	۰/۹۶۰	۰/۹۴۶	۰/۹۴۶			
	X_5	-۰/۰۳۸۱	-۰/۰۳۸۶	-۰/۰۵۲۹	۱/۵۷۹	۱/۴۲۹	۱/۵۵۸	۰/۹۷۱	۰/۹۵۶	۰/۹۵۳	۰/۰۳۲	۰/۰۴۳	۰/۰۴۶
MAR(۰/۳،۰/۴)	عرض از مبدا	-۰/۰۴۱۲	۰/۰۷۴۴	۰/۱۲۸۴	۱/۴۰۳	۱/۲۳۶	۱/۴۶۱	۰/۹۶۸	۰/۹۶۹	۰/۹۲۷			
	X_1	-۰/۰۰۲۹	-۰/۰۱۳۳	-۰/۰۳۹۱	۱/۲۷۵	۱/۱۸۸	۱/۴۶۸	۰/۹۶۹	۰/۹۵۶	۰/۹۴۷			
	X_2	-۰/۰۰۵۲	۰/۰۱۴۶	-۰/۰۵۲۸	۱/۴۴۹	۱/۳۳۰	۱/۷۰۸	۰/۹۶۴	۰/۹۴۵	۰/۹۵۷			
	X_3	۰/۰۲۲۶	-۰/۱۶۴۶	۰/۰۲۰۲	۱/۶۷۸	۱/۴۲۶	۱/۴۸۵	۰/۹۶۸	۰/۹۸۴	۰/۹۴۴			
	X_4	-۰/۰۱۴۵	۰/۰۱۳۹	-۰/۰۴۵۴	۱/۲۶۰	۱/۱۶۷	۱/۴۴۲	۰/۹۶۵	۰/۹۴۵	۰/۹۵۲			
	X_5	-۰/۰۲۷۵	-۰/۰۲۶۲	-۰/۰۵۱۸	۱/۵۸۳	۱/۴۲۳	۱/۷۸۵	۰/۹۷۴	۰/۹۵۵	۰/۹۴۷	۰/۰۲۷	۰/۰۵۲	۰/۰۵۳
MCAR(۰/۴)	عرض از مبدا	-۰/۰۵۴۰	۰/۰۵۳۲	-۰/۰۵۷۸	۱/۴۱۴	۱/۲۳۶	۱/۵۶۶	۰/۹۷۴	۰/۹۶۹	۰/۹۵۷			
	X_1	-۰/۰۰۸۶	-۰/۰۱۹۰	-۰/۰۰۹۹	۱/۲۷۸	۱/۱۸۸	۱/۵۶۱	۰/۹۶۶	۰/۹۵۵	۰/۹۵۱			
	X_2	-۰/۰۱۱۴	۰/۰۰۹۲	-۰/۰۲۷۸	۱/۴۶۲	۱/۳۳۳	۱/۷۶۵	۰/۹۷۵	۰/۹۴۹	۰/۹۶۴			
	X_3	۰/۰۴۱۰	-۰/۱۲۹۸	۰/۰۲۶۶	۱/۷۰۲	۱/۴۲۹	۱/۵۸۴	۰/۹۷۰	۰/۹۸۹	۰/۹۵۷			
	X_4	-۰/۰۰۰۷	۰/۰۲۹۲	-۰/۰۳۷۱	۱/۲۶۱	۱/۱۶۴	۱/۵۴۰	۰/۹۶۵	۰/۹۴۵	۰/۹۴۹			
	X_5	-۰/۰۲۷۵	-۰/۰۲۵۸	-۰/۰۲۹۷	۱/۵۹۱	۱/۴۲۲	۱/۹۱۱	۰/۹۷۱	۰/۹۵۰	۰/۹۴۸	۰/۰۳۲	۰/۰۵۰	۰/۰۵۳

توضیح: بیش‌ترین اریبی و کارایی برای هر برآورد ارایه‌شده به‌صورت پررنگ نشان داده شده است. MCAR = گم‌شده‌ی کاملاً تصادفی؛ MAR = گم‌شده‌ی تصادفی؛ CC = مورد کامل؛ EM = ماکسیمم‌سازی امید ریاضی؛ MI = جانهی چندگانه.

الف. نرخ عدم پذیرش به‌صورت درصد دفعاتی که فرض صفر درباره‌ی ضریب رگرسیونی جامعه‌ی X_5 در میان ۱۰۰۰ نمونه رد شده است تعریف می‌شود؛ این نرخ باید با مقدار اسمی $\alpha = ۰/۰۵$ مقایسه شود. بدترین نرخ عدم پذیرش به‌صورت پررنگ نشان داده شده است.

جدول ۳- اریبی، کارایی، پوشش و نرخ عدم پذیرش روش‌های EM، MI و CC برای $n = ۲۰۰$ و p برنولی= ۰/۲

		اریبی			کارایی			پوشش سطح واقعی= ۹۵٪			نرخ عدم پذیرش[الف] مقدار اسمی $\alpha = ۰/۰۵$		
		EM	MI	CC	EM	MI	CC	EM	MI	CC	EM	MI	CC
MAR(۰/۱،۰/۲)	عرض از مبدا	-۰/۰۷۵۴	۰/۰۰۶۶	۰/۰۵۱۷	۲/۰۱۸	۱/۷۵۸	۱/۸۵۷	۰/۹۷۲	۰/۹۶۷	۰/۹۳۹			
	X_1	۰/۰۰۰۴	-۰/۰۰۴۸	-۰/۰۱۱۰	۱/۹۸۲	۱/۷۱۳	۱/۸۵۹	۰/۹۷۷	۰/۹۵۵	۰/۹۵۲			
	X_2	-۰/۰۴۰۵	-۰/۰۳۰۰	-۰/۰۶۵۴	۲/۳۰۸	۱/۹۳۶	۲/۵۰۰	۰/۹۷۹	۰/۹۵۴	۰/۹۳۶			
	X_3	۰/۰۴۵۶	-۰/۰۷۹۷	۰/۰۰۲۷	۲/۲۸۳	۱/۹۰۶	۱/۸۷۹	۰/۹۸۰	۰/۹۸۳	۰/۹۵۲			
	X_4	-۰/۰۳۵۶	-۰/۰۱۹۹	-۰/۰۵۳۷	۱/۹۵۵	۱/۶۹۶	۱/۸۲۹	۰/۹۷۶	۰/۹۵۷	۰/۹۵۱			
	X_5	-۰/۰۵۶۳	-۰/۰۵۴۳	-۰/۰۸۵۶	۳/۱۵۳	۲/۰۶۴	۲/۴۲۸	۰/۹۹۱	۰/۹۶۸	۰/۹۶۴	۰/۰۱۰	۰/۰۳۶	۰/۰۳۵
MAR(۰/۳،۰/۴)	عرض از مبدا	-۰/۰۷۵۹	۰/۰۵۹۲	۰/۰۳۶۷	۲/۲۲۲	۱/۷۸۸	۲/۲۴۶	۰/۹۷۶	۰/۹۶۷	۰/۹۳۰			
	X_1	-۰/۰۱۷۰	-۰/۰۲۴۱	-۰/۰۲۹۹	۲/۰۱۴	۱/۷۲۶	۲/۲۶۰	۰/۹۷۷	۰/۹۵۲	۰/۹۴۶			
	X_2	-۰/۰۷۳۷	-۰/۰۵۵۱	-۰/۲۰۲۵	۲/۳۶۹	۱/۹۶۰	۳/۳۴۱	۰/۹۷۸	۰/۹۴۸	۰/۹۶۹			
	X_3	۰/۰۲۸۱	-۰/۱۷۷۵	۰/۰۸۷۳	۲/۶۸۳	۲/۰۶۳	۲/۲۷۲	۰/۹۸۲	۰/۹۹۵	۰/۹۶۶			
	X_4	-۰/۰۲۴۹	۰/۰۰۳۴	-۰/۰۵۷۹	۱/۹۸۴	۱/۷۰۰	۲/۱۲۰	۰/۹۶۹	۰/۹۴۱	۰/۹۴۳			
	X_5	-۰/۰۶۸۹	-۰/۰۶۶۰	-۰/۱۶۴۷	۲/۶۸۵	۲/۰۸۱	۳/۶۴۰	۰/۹۸۹	۰/۹۶۷	۰/۹۶۰	۰/۰۱۰	۰/۰۳۶	۰/۰۳۹
MCAR(۰/۴)	عرض از مبدا	-۰/۱۰۵۴	۰/۰۱۰۷	-۰/۱۹۲۰	۲/۲۵۵	۱/۸۱۳	۲/۵۱۹	۰/۹۸۲	۰/۹۷۸	۰/۹۵۶			
	X_1	-۰/۰۱۵۸	-۰/۰۲۲۵	-۰/۰۲۰۱	۲/۰۳۸	۱/۷۲۳	۲/۳۴۱	۰/۹۷۹	۰/۹۶۲	۰/۹۵۶			
	X_2	-۰/۰۵۷۱	-۰/۰۳۳۵	-۰/۱۵۴۳	۲/۵۴۳	۱/۹۵۵	۳/۵۰۴	۰/۹۸۲	۰/۹۵۲	۰/۹۶۷			
	X_3	۰/۰۶۷۷	-۰/۱۰۹۷	۰/۱۰۴۸	۲/۷۱۸	۲/۰۷۹	۲/۵۳۲	۰/۹۸۴	۰/۹۹۸	۰/۹۶۳			
	X_4	-۰/۰۱۰۳	۰/۰۱۹۶	-۰/۰۶۳۸	۱/۹۹۲	۱/۶۹۲	۲/۲۹۳	۰/۹۷۱	۰/۹۴۸	۰/۹۶۳			
	X_5	-۰/۰۵۵۰	-۰/۰۵۳۷	-۰/۲۰۵۰	۲/۷۲۷	۲/۰۷۶	۴/۵۶۰	۰/۹۸۸	۰/۹۷۰	۰/۹۶۱	۰/۰۱۲	۰/۰۳۲	۰/۰۳۶

توضیح: بیش‌ترین اریبی و کارایی برای هر براورد ارائه‌شده به‌صورت پررنگ نشان داده شده است. MCAR = گم‌شده‌ی کاملاً تصادفی؛ MAR = گم‌شده‌ی تصادفی؛ CC = مورد کامل؛ EM = ماکسیمم‌سازی امید ریاضی؛ MI = جانهی چندگانه.

الف. نرخ عدم پذیرش به‌صورت درصد دفعاتی که فرض صفر درباره‌ی ضریب رگرسیونی جامعه‌ی X_5 در میان ۱۰۰۰ نمونه رد شده است تعریف می‌شود. این نرخ باید با مقدار اسمی $\alpha = ۰/۰۵$ مقایسه شود. بدترین نرخ عدم پذیرش به‌صورت پررنگ نشان داده شده است.

فراوانی به دست می‌دهد. بدترین عملکرد CC روی اریبی در سه مکانیزم داده‌های گم‌شده و دو اندازه‌ی نمونه نفوذ کرده است. EM در مورد اریبی اندکی بهتر از MI عمل کرده است. براوردهای MI ضریب‌های رگرسیونی X_4 و X_5 شامل اریبی کم‌تری از EMها در الگوهای MAR بوده و براوردهای ضریب‌های رگرسیونی X_2 و X_5 دارای اریبی کم‌تری از EMها در الگوی MCAR بوده‌اند. برای ۶ حالت نشان داده شده در جدول‌های ۲ و ۳، MI اریب‌ترین براورد را برای X_3 و کم‌ترین براورد اریبی را برای X_5 تولید کرده است. عرض از مبدأ Y توسط MI درست‌تر از EM براورد شد.

کارایی: از آن‌جا که فاصله‌ی پهن‌تر حاکی از رهیافتی با کارایی کم‌تر بوده، پهن‌ترین و در واقع بدترین فاصله‌ی ۹۵% مشخص شده است. MI به‌صورت یکنواخت بهترین رهیافت از لحاظ کارایی بود. CC بدترین رهیافت تحت (۰/۳,۰/۴) MAR و MCAR بوده و EM بدترین رهیافت تحت (۰/۱,۰/۲) MAR، بدون در نظر گرفتن اندازه‌ی نمونه بوده‌اند. به‌علاوه، EM در براورد ضریب X_3 به‌طور یکنواخت رهیافتی با کم‌ترین کارایی بوده است.

پوشش: هر سه رهیافت نرخ‌های پوشش مشابهی ارایه دادند. بنا بر [۲۳]، پوشش کم‌تر از ۹۰% نشان‌دهنده‌ی یک پوشش جداً سطح پایین است، زیرا ۹۰% برابر است با دو برابر نرخ خطای اسمی (۵%). با این توضیح، هر سه رهیافت به‌طور مساوی فواصل اطمینان خوبی برای همه‌ی متغیرهای کمکی تحت همه‌ی حالت‌ها ارایه دادند.

نرخ عدم پذیرش: این معیار تنها برای X_5 دارای کاربرد است زیرا این متغیر کمکی نامرتبط با X_1 تا X_4 و Y تولید شده و لذا ضریب واقعی آن در جامعه برابر صفر است. جدول‌های ۲ و ۳ نشان می‌دهند که EM به‌طور یکنواخت بدترین رهیافت است زیرا فرض صفر را محتاطانه‌تر از هر دو رهیافت MI و CC رد می‌کند. نرخ عدم پذیرش محتاطانه توسط EM برای اندازه‌ی نمونه‌ی ۲۰۰ بدتر از ۴۰۰ بود. زمانی که $n = ۲۰۰$، نرخ عدم پذیرش EM به کمی ۰/۰۱ است، یعنی تنها ۲۰% سطح اسمی α (۰/۰۵). قابل درک است که همه‌ی رهیافت‌ها بدترین نرخ عدم پذیرش را برای $n = ۲۰۰$ و الگوی داده‌های گم‌شده‌ی (۰/۱,۰/۲) MAR ارایه داده‌اند. هنگامی که اندازه‌ی نمونه به $n = ۴۰۰$ افزایش پیدا کرد همه‌ی رهیافت‌ها نرخ‌های عدم پذیرش بهتری ارایه دادند. در $n = ۴۰۰$ بهترین نرخ عدم پذیرش تحت (۰/۳,۰/۴) MAR، سپس MCAR و پس

از آن (۰/۱,۰/۲) MAR به دست آمد، هر چند سه نرخ عدم پذیرش برای هر رهیافت بیش از ۰/۰۱۲ اختلاف نداشتند.

اثر مکانیزم داده‌های گم‌شده، نرخ‌های گم‌شدگی ، و اندازه‌های نمونه

MAR، نرخ‌های گم‌شدگی، و اندازه‌های نمونه. با در نظر گرفتن مکانیزم‌های داده‌های گم‌شده‌ی MAR به تنهایی و اندازه‌ی نمونه‌ی بزرگ‌تر (یعنی، $n = ۴۰۰$)، نتیجه‌های جدول ۲ نشان می‌دهند که اریبی براوردهای حاصل از EM، MI و CC به‌طور کلی زمانی که نرخ گم‌شدگی زیاد بود (۰/۱,۰/۲) کم‌تر از زمانی بود که نرخ گم‌شدگی کم (۰/۳,۰/۴) بود. مورد استثناء $X_۵$ بود که زمانی که نرخ گم‌شدگی زیادتر بود اریبی آن کم‌تر بود. همچنین، براوردهای EM برای $X_۱$ و $X_۲$ زمانی که نرخ‌های گم‌شدگی بیش‌تر بود اریبی کم‌تری داشتند. EM و CC زمانی که نرخ گم‌شدگی کم بود (۰/۱,۰/۲) هر دو کارامدتر از زمانی بودند که این نرخ زیاد (۰/۳,۰/۴) بود. اختلاف کارایی بین دو نرخ گم‌شدگی برای CC فاحش‌تر از EM است. MI تقریباً برای هر دو نرخ گم‌شدگی به یک اندازه کارامد است.

همان‌طور که در بالا گفته شد هر سه رهیافت پوشش قابل قبولی از پارامترهای جامعه به دست دادند. به‌علاوه، به‌نظر می‌رسید که پوشش مستقل از نرخ گم‌شدگی است. نرخ‌های عدم پذیرش توسط MI و CC زمانی که نرخ گم‌شدگی زیاد بود، به سطح اسمی ۰/۰۵ نزدیک‌تر از زمانی بودند که نرخ گم‌شدگی کم بود. این روند برای EM، که بدترین و محافظه‌کارانه‌ترین نرخ عدم پذیرش را به دست می‌داد، معکوس می‌شد. برای الگوی داده‌های گم‌شده‌ی MAR و اندازه‌ی نمونه‌ی کوچک (یعنی، $n = ۲۰۰$)، نتیجه‌های نشان داده‌شده در جدول ۳ مشابه یافته‌های نمونه‌ی بزرگ‌تر روی هر ۴ معیار است. چند استثناء در زیر ذکر شده است:

- اریبی در $X_۳$ و $X_۴$ برای EM زمانی که نرخ گم‌شدگی بالاتر است کم‌تر از زمانی است که نرخ گم‌شدگی کم‌تر است.
- اریبی در عرض از مبدأ Y برای CC در نرخ گم‌شدگی بالا کم‌تر از نرخ گم‌شدگی پایین است.

- کارایی EM در براورد ضریب‌های X_5 با نرخ گم‌شدگی پایین‌تر کم‌تر از زمانی است که نرخ گم‌شدگی زیادتر است.
- نرخ‌های عدم پذیرش توسط سه رهیافت برای هر دو نرخ گم‌شدگی تقریباً یکسان است.

اگر نرخ گم‌شدگی ثابت نگاه داشته شود، نتیجه‌های حاصل از نمونه‌ی بزرگ از لحاظ کارایی و نرخ عدم پذیرش بهتر از نمونه‌ی کوچک است؛ MAR (۰/۱,۰/۲) و MAR (۰/۳,۰/۴) را از جدول‌های ۲ و ۳ مقایسه کنید. پوشش توسط سه رهیافت مستقل از اندازه‌ی نمونه به نظر می‌رسد. اریبی تولید شده توسط MI و CC برای عرض از مبدأ Y با نمونه‌های کوچک کم‌تر از نمونه‌های بزرگ است. میزان اریبی در متغیرهای کمکی دیگر مستقل از اندازه‌ی نمونه برای MI و CC به نظر می‌رسد.

MCAR و اندازه‌های نمونه. تحت مکانیزم MCAR و برای هر دو اندازه‌ی نمونه‌های بزرگ و کوچک، همان‌طور که در جدول‌های ۲ و ۳ نشان داده شده است، EM برای بیش‌تر متغیرهای کمکی اریبی کم‌تری نسبت به MI و CC داشته است. MI و CC تنها زمانی که اندازه‌ی نمونه بزرگ بود نرخ‌های عدم پذیرش نزدیک‌تر به سطح اسمی α (۰/۰۵) ارایه دادند. زمانی که اندازه‌ی نمونه به ۲۰۰ کاهش یافت نرخ‌های عدم پذیرش توسط هر سه رهیافت بدتر شد. نرخ عدم پذیرش EM فارغ از اندازه‌ی نمونه از همه بدتر و محافظه‌کارانه‌تر بود.

مشابه یافته‌هایی که تحت MAR به دست آمد، نتیجه‌های نمونه‌های بزرگ بر مبنای MCAR بر حسب کارایی و نرخ عدم پذیرش بهتر از نتیجه‌های نمونه‌های کوچک بود. پوشش توسط هر سه رهیافت مستقل از اندازه‌ی نمونه به نظر می‌رسید. همان‌طور که قبلاً اشاره شد، هر سه رهیافت پوشش قابل قبولی برای هر دو $n=۲۰۰$ و $n=۴۰۰$ به دست دادند. اریبی در مجموع با نمونه‌های بزرگ کم‌تر از نمونه‌های کوچک بود، به استثناء اریبی تولید شده توسط MI و CC برای عرض از مبدأ Y.

نتیجه‌گیری:

بر اساس نتیجه‌های به دست آمده می‌توان به نتیجه‌گیری‌های زیر رسید:

۱- به‌نظر می‌رسد پوشش پارامتر واقعی توسط سه رهیافت مستقل از اندازه‌ی نمونه، مکانیزم داده‌های گم‌شده، و نرخ‌های گم‌شدگی است. هرسه رهیافت به‌طور یکنواخت پوشش قابل قبولی تحت همه‌ی شرایط به دست دادند که هیچ یک کم‌تر از ۹۰% نبود.

۲- MI از لحاظ کارایی تحت هر دو مکانیزم MAR و MCAR برای هر دو اندازه‌ی نمونه به‌صورت سازگار بهتر از EM عمل کرده است. زمانی که نرخ گم‌شدگی پایین بود، کارایی EM بدتر از CC بوده است. تحت همه‌ی شرایط دیگر، براوردهای CC کم‌ترین کارایی را داشته‌اند.

۳- MI از نظر نرخ عدم پذیرش تحت هر دو مکانیزم MAR و MCAR برای هر دو اندازه‌ی نمونه به‌صورت سازگار بهتر از EM عمل کرده است. نرخ عدم پذیرش EM زمانی که اندازه‌ی نمونه کوچک بود، فارغ از مکانیزم گم‌شدگی به شدت محافظه‌کارانه بود. نرخ عدم پذیرش MI خیلی نزدیک به CC بود. هر دو نرخ عدم پذیرش زمانی که اندازه‌ی نمونه بزرگ بود به نرخ اسمی ۰/۰۵ نزدیک‌تر از زمانی بودند که اندازه‌ی نمونه کوچک بود.

۴- از نظر اریبی، رهیافت CC بدترین اریبی را با بیش‌ترین فراوانی داشته است. EM زمانی که اندازه‌ی نمونه بزرگ ($n = ۴۰۰$) و مکانیزم گم‌شدگی MCAR بود بهتر از MI عمل کرد. برای شش حالت تحت مطالعه، MI اریب‌ترین براورد را برای $X_۳$ و براوردی با کم‌ترین اریبی را برای $X_۵$ به دست داد. عرض از مبدأ Y توسط MI صحیح‌تر از EM براورد شد. میزان اریبی تولید شده توسط MI و CC برای عرض از مبدأ Y به اندازه‌ی نمونه بستگی داشت به این شکل که در نمونه‌های کوچک فارغ از مکانیزم داده‌های گم‌شده، کم‌تر بود. اریبی در دیگر متغیرهای کمکی تولید شده توسط MI و CC به اندازه‌ی نمونه و مکانیزم‌های داده‌های گم‌شده بستگی داشت.

۵- نتیجه‌های حاصل از نمونه‌ی بزرگ از لحاظ کارایی و نرخ عدم پذیرش، فارغ از مکانیزم گم‌شدگی داده‌ها و نرخ‌های گم‌شدگی بهتر از نتایج نمونه‌ی کوچک بود. از لحاظ اریبی و پوشش، عملکرد EM و MI هر دو به‌طور کلی مستقل از اندازه‌ی نمونه و مکانیزم داده‌های گم‌شده بود.

۶- دو الگـوی گـم‌شـدگی (۰/۴) MCAR و (۰/۳,۰/۴) MAR هـیچ یـک از دو عملکرد EM و MI را بـه‌صـورت متفـاوت تحـت تـأثیر قـرار ندادنـد. ایـن پدیـده می‌تواند توسط این واقعیت که نـرخ‌هـای گـم‌شـدگی در هـر دو الگـو مـشابه بـوده، MCAR حالت خاصی از MAR است و در هر دو الگوریتم قاعده‌ی بیز بـه کـار رفته، توضیح داده شود.

۷- زمانی کـه هـر چهـار معیـار همزمـان در نظـر گرفتـه شـوند، EM و MI در حالـت اندازه‌ی نمونه‌ی بزرگ ($n = ۴۰۰$) و نرخ گم‌شدگی پایین در (۰/۱,۰/۲) MAR بهترین عملکرد را داشتند. از آن‌جا که هر دو رهیافـت بـر مبنـای درسـتنمایی‌انـد، این امر مورد انتظار بود.

یافته‌ها به‌طور کلی MI را بر EM ترجیح می‌دادند. MI از لحاظ کارایی و نـرخ عـدم پذیرش بـه‌طـور سـازگار بهتـر از EM عمـل مـی‌کـرد. پوشـش MI در سـطح ۹۵% بـه‌طـور یکنواخت از EM بـه ۹۵% نزدیـک‌تـر بـود، بـه اسـتثنای متغیـر کمکـی $X_۳$ بـا داده‌هـای گم‌شده. از نظر اریبی EM کمی بهتر از MI عمل کرد. براورد MI برای $X_۳$ به‌طور سازگار اریب‌ترین براورد تحت همه‌ی حالت‌های مورد بررسی بود. EM زمـانی کـه انـدازه‌ی نمونـه بزرگ ($n = ۴۰۰$) و پارامتر برنولی (p) کوچک (۰/۲) بود براوردهایی با اریبی کم‌تـر از MI بـه دسـت داد. زمـانی کـه انـدازه‌ی نمونـه بـه ۲۰۰ کـاهش پیـدا کـرد، براوردهـای MI برای $X_۵$ (متغیر دوجمله‌ای) و عـرض از مبـدأ Y فـارغ از مکـانیزم داده‌هـای گـم‌شـده یـا پارامتر برنولی کم‌تر از براوردهای EM اریبی داشتند. از آن‌جا که هر دو الگوریتم به‌صورت مجانبی سازگارند، اریبی موضوعی جدی نیست مگر زمانی که اندازه‌ی نمونه کوچک باشد.

همـان‌گونـه کـه انتظـار مـی‌رفـت، هـر دو رهیافـت زمـانی کـه انـدازه‌ی نمونـه بـزرگ ($n = ۴۰۰$) و نرخ گم‌شدگی پایین (۰/۱,۰/۲) تحـت MAR بـود، نتیجـه‌هـای بهتـری بـه دست دادند. زمانی که نرخ گم‌شدگی بالا و انـدازه‌ی نمونـه کوچـک بـود، بـرای EM و MI میسر نبود که اطلاعات کافی از داده‌های مشاهده‌شده به دست آورند تا داده‌های گـم‌شـده را پـیش‌بینـی کننـد. در نتیجـه، هـر دو رهیافـت، بـا کـارایی مجـانبی، بـا $n = ۲۰۰$ عملکـرد ضعیفی داشتند.

۷- بحث

یافته‌های مبتنی بر پارامتر دوجمله‌ای ۰/۴ بسیار شبیه نتیجه‌های حاصل از $p = ۰/۲$ بودند؛ این یافته‌ها هفت نتیجه‌گیری ارایه‌شده در بالا را تقویت می‌کنند. کارایی و پوشش تحت تأثیر پارامترهای دوجمله‌ای قرار نداشتند. پوشش به‌دست آمده توسط سه رهیافت از دو پارامتر دو جمله‌ای به‌طور مساوی قابل قبول بود. ولی هنوز تفاوت‌های کوچکی باقی می‌ماند:

- با $n = ۴۰۰$ بیش‌تر اریبی‌ها با $p = ۰/۴$ کمی بزرگ‌تر از $p = ۰/۲$ برای همه‌ی مکانیزم‌های داده‌های گم‌شده بودند.
- نرخ عدم پذیرش EM که بدترین در میان سه رهیافت است، با $p = ۰/۴$ به سطح اسمی ۰/۰۵ نزدیک‌تر از $p = ۰/۲$ برای دو اندازه‌ی نمونه بود.

دشواری‌های محاسباتی

چندین مشکل محاسباتی در مطالعه‌ی حاضر رخ داده که بیش‌ترین آن‌ها در الگوریتم EM بوده است. اول، زمانی که $n = ۲۰۰$ بود، برای برنامه‌ی S-Plus محاسبه‌ی خطاهای استاندارد در اجرای EM دشوار بود به خصوص برای $X_۳$ که شامل داده‌های گم‌شده بود. دلیل این امر آن بود که اندازه‌ی نمونه در مقایسه با تعداد پارامترها ($۲×۲×۲×۲×۲-۱=۳۱$) در یک مدل لوژستیک با ۵ متغیر دو حالتی، خیلی کوچک بود. به‌علاوه، نسبت Yهای برابر ۱ (ناراضی از کارت اعتباری) در جامعه تنها ۰/۱۴۸۱۱۲ بود. در نتیجه مشاهداتی با Y برابر ۱ در نمونه‌هایی به اندازه‌ی ۲۰۰ کم بود. باور بر این است که این عوامل در دشواری EM برای محاسبه‌ی خطاهای استاندارد سهم داشته‌اند.

دومین دشواری رهیافت EM پیچیدگی محاسباتی آن بود. محققان علاقه‌مند به استفاده از روش وزنی EM باید کدهای برنامه‌نویسی خاص خود را در MATLAB، S-Plus یا زبان R خود رأساً بنویسند زیرا این الگوریتم EM در یک بسته‌ی آماری برای مقاصد عام مانند SPSS یا SAS گنجانده نشده است. نرم افزارهای رایگانی مانند LEM یا نسخه‌ی جدیدتر آن Latent Gold 4.0 می‌توانند براوردهای ML پارامترهای رگرسیونی لوژستیک را به‌سادگی ارایه دهند. هرچند از وب سایت آن در ۲۵ فوریه ۲۰۰۷

(http://www.statisticalinnovations.com/products/latentgold_v4.html)
مشخص نیست که آیا این برنامه‌ها می‌توانند برای اجرای روش وزن‌های ابراهیم [۱۰] مطابقت داده شوند. مشکل سوم رهیافت EM خاص بودن محاسباتی آن است. اجرای روش ابراهیم برای مدل‌هایی با متغیرهای کمکی گسسته و داده‌های گم‌شده‌ای که تنها به برامدها و داده‌های مشاهده‌شده، یعنی مکانیزم MAR ([۹]) وابسته‌اند، آسان است. اگر مدلی شامل متغیرهای کمکی پیوسته، آمیزه‌ای از متغیرهای کمکی پیوسته و گسسته، و/یا نرخ گم‌شدگی بالا باشد، الگوریتم EM ساخته‌شده توسط هورتون و لِرد باید برای مطابقت با مدل تصحیح شود. این کار به هیچ وجه آسان یا ساده نیست. سرانجام، اگر درصد داده‌های گم‌شده بالا باشد نرخ همگرایی می‌تواند به شدت کند باشد ([۲] و [۱۳]). این امر به این دلیل است که نرخ همگرایی برای روش‌های تکرار مبتنی بر درستنمایی مانند EM و روش وزن‌های ابراهیم با درصد داده‌های کاملاً مشاهده‌شده در یک مجموعه‌ی داده‌ها متناسب است ([۱۳]، ص ۱۳۰).

ولی درباره‌ی امکان تعمیم روش وزن‌های EM به متغیرهای کمکی پیوسته، هورتون و لِرد ([۹]، صص۴۳-۴۴) اظهار داشته‌اند که چنین فنونی، با یا بدون یک مدل پارامتری برای متغیرهای کمکی پیوسته موجود است؛ ولی از نظر محاسباتی پیچیده‌اند. همچنین گسترش رهیافت EM به مدل‌هایی با اثر متقابل از نظر محاسباتی پیچیده است، اگرچه غیر ممکن نیست [۲۳]. لذا کاربرد روش وزن‌های EM محدود است.

این پدیده‌ها به یک مزیت واضح رهیافت MI اشاره می‌کنند؛ یعنی محققان تنها یک برنامه‌ی S-Plus برای MI نیاز دارند. زمانی که چند مجموعه‌ی داده‌های جانهی‌شده ساخته می‌شود، می‌توان هر مدل آماری را برای این مجموعه‌ی داده‌های جانهی‌شده در هر نرم‌افزار آماری به کار برد. در مقابل، رهیافت EM به یک برنامه‌ی مخصوص که خاص اجرای مدل رگرسیونی لوژستیک یا یک مدل انتخابی نوشته شده باشد نیاز دارد. برنامه‌های رایانه‌ای مدل- ویژه به‌طور کلی ترجیح داده نمی‌شوند مگر این که پژوهش‌گران زمان و منابع کافی در اختیار داشته باشند. به‌علاوه، مدل جانهی و مدل تحلیل برای MI لازم نیست به مدل‌هایی با متغیرهای کمکی دو حالتی و/یا یک متغیر برامد دو حالتی محدود شود. این مدل‌ها می‌توانند شامل اثر متقابل بین متغیرهای کمکی نیز باشند. در مجموع، MI برای هر مدل خطی تعمیم‌یافته‌ی مفروض برای پژوهش‌های آموزشی و روان‌شناختی کاربرد دارد. لذا، اجرای MI نسبتاً ساده است. سرانجام، MI در مطالعه‌ی

حاضر و مطالعه‌های دیگر عملکرد خوبی داشته است [۷]. حتی در وضعیت‌هایی که MI بهتر از الگوریتم‌های مدل- ویژه یا مشکل- ویژه عمل نمی‌کند، هنوز هم رهیافتی راحت و کارایی برای برخورد با داده‌های گم‌شده در مجموعه‌ی داده‌های تجربی محسوب می‌شود.

مدل جانهی در MI

کلید موفقیت رهیافت MI تعیین مدل جانهی است. MI فرض می‌کند که مدل جانهی با مدل تحلیل که متعاقباً برای تحلیل داده‌ها به کار خواهد رفت یکسان است. هر چند در عمل نیازی نیست که این دو مدل یکسان باشند. بنا بر [۲۱]، مدل جانهی باید شامل پیشگوهایی باشد که دارای اهمیت ذاتی باشند گم‌شدگی و متغیرهایی با داده‌های گم‌شده را به‌خوبی پیشگویی کنند، و ویژگی‌های به خصوص طرح نمونه (به‌عنوان مثال، آمارگیری‌های احتمالاتی) را بازتاب دهند. مدل جانهی نیازی نیست که از نظر مفهومی معنی داشته باشد. در عمل، این مدل نوعاً شامل متغیرهای ذاتاً مورد علاقه و همچنین پیشگوهایی است که منعکس‌کننده‌ی گم‌شدگی پنداشته می‌شوند. شواهد تجربی نشان می‌دهند که مدل جانهی حتی اگر درست تعیین نشده باشد [۲۱] می‌تواند در کاهش اریبی مفید باشد.

تعداد جانهی‌ها

در مورد سئوال «چند جانهی برای MI مورد نیاز است؟» رابین ([۱۹]، ص۱۱۴) نشان می‌دهد که کارایی یک براورد پارامتر بر اساس m جانهی برابر است با $\left(1+\frac{\gamma}{m}\right)^{-1}$ که در آن γ نرخ گم‌شدگی اطلاعات است. نرخ اطلاعات گم‌شده به‌صورت افزایش واریانس برارود پارامتر ناشی از داده‌های گم‌شده در یک نمونه‌ی بزرگ تعریف می‌شود؛ که می‌تواند از نرخ مشاهده‌های گم‌شده کم‌تر یا بیش‌تر باشد. برای نرخ‌های پایین اطلاعات گم‌شده (۲۰% یا کم‌تر)، براوردی مبتنی بر $m=3$ جانهی دارای کارایی برابر حد اقل ۹۴% است. برای نرخ‌های بالاتر اطلاعات گم‌شده، جانهی‌های اضافی مورد نیاز است. شیوه‌ی MI در SAS نسخه‌ی ۸ و ۹ پیش‌فرض تعداد جانهی‌ها را برابر ۵ قرار می‌دهد (SAS Institute، 2004، [۲۰]). به‌طور کلی، توصیه می‌شود که مقادیر گم‌شده پیش از ادغام نتایج ۵ بار جانهی شوند [۱۶].

نگرانی‌های دیگر موجود در MI

علی‌رغم نقطه‌های قوت آن، نگرانی‌های بسیاری در مورد MI در نوشتگان مستندسازی شده است. اولین آن‌ها این است که هر کاربرد MI مقادیر جانهی‌شده و آماره‌های مرتبط کمی متفاوت تولید می‌کند. در نتیجه، زمانی که MI برای برخورد با داده‌های گم‌شده مورد استفاده قرار می‌گیرد، همیشه نمی‌توان نتیجه‌ها را تکرار کرد. با مساوی نگه داشتن بقیه‌ی موارد، این مسئله با زیاد شدن مقدار داده‌های گم‌شده اهمیت بیش‌تری می‌یابد. دوم این‌که، آلیسون ([۱]) مشکلات MI را زمانی که براورد اثر متقابل مد نظر است مورد بحث قرار می‌دهد. سومین نگرانی محدودیت دسترسی به نرم‌افزار MI در نرم‌افزارهای آماری معمول است. در کنار دو شیوه‌ی (MI، MIANALYZE) مورد اجرا در SAS نسخه‌ی ۸/۲ و بالاتر، نرم افزار رایگان MI، NORM که توسط شیفر ([۲۱]) نوشته شده است نیز می‌تواند برای اجرای MI تحت فرض نرمال چندمتغیره به کار گرفته شود. این نرم افزار به همراه سه بخش اضافی: CAT، MTX و PAN از طریق http://www.Stat.Psu.edu/~jls/misotwa.html در دسترس است. CAT، MI را برای داده‌های رسته‌ای تحت فرض توزیع لگ‌خطی (Log-Linear) اجرا می‌کند، MIX، MI را برای آمیزه‌ای از داده‌های پیوسته و رسته‌ای تحت فرض توزیع مکانی عام اجرا می‌کند، و PAN، MI را برای داده‌های پانلی یا داده‌های خوشه‌ای تحت یک توزیع چند متغیره‌ی خطی با اثرهای آمیخته اجرا می‌کند. با این‌که نرم‌افزار شیفر به کارگیری MI را در انواع گوناگونی از حالت‌های داده‌های گم‌شده میسر می‌سازد، پژوهشگران ممکن است هنوز این برنامه‌ها را کم‌تر از شیوه‌های MI و MIANALYZE در SAS کاربردوست بیابند.

پیامدهایی برای پژوهشگران

مطالعه‌ی حاضر در پی گسترش کار هورتون و لِرد در [۹] با گنجاندن رهیافت MI برای برخورد با متغیرهای کمکی دو حالتی گم‌شده در مدل‌های رگرسیونی لوژستیک بوده است. این مطالعه به تمجید از کار گراهام و شیفر، [۷] با بررسی عملکرد MI برای هر دو مجموعه‌ی داده‌های بزرگ و کوچک پرداخته است. به ویژه، روش وزن‌های EM و MI از لحاظ اریبی، کارایی، پوشش و نرخ عدم پذیرش مقایسه شده‌اند. یافته‌های ما نشان می‌دهند که زمانی که داده‌های گم‌شده به‌طور تصادفی رخ می‌دهند (یعنی MAR)، MI

می‌تواند یک رهیافت ماندنی برای برخورد با متغیرهای کمکی دو حالتی گم‌شده در رگرسیون لوژستیک باشد. روش وزن‌های EM، که نیازمند یک برنامه‌ی رایانه‌ای است که ویژه‌ی هر مدل انتخابی پژوهشگر نوشته شده باشد، نه کاربردی است و نه برتر از MI تلقی می‌شود.

با کدبندی دو حالتی متغیر وابسته و هر ۴ متغیر کمکی، یافته‌های مطالعه‌ی حاضر با معیارهای روان‌شناختی و آموزشی مناسبت دارد. به ویژه، مدل‌های لوژستیک به وفور برای ربط دادن اندازه‌های برامد دوجمله‌ای (یعنی، قبول یا رد، صحیح یا ناصحیح) با متغیرهای کمکی آن‌ها، اعم از رسته‌ای (نوع مدرسه)، پیوسته (سن)، یا آمیزه‌ای از هر دو مورد استفاده قرار می‌گیرند. MI حتی با وجود اثر متقابل می‌تواند برای برخورد مناسب با داده‌های گم‌شده در متغیرهای کمکی به کار رود. هورتون و لِرد [۹] گسترشی از روش وزن‌های EM را برای پرداختن به خطاهای بدرده‌بندی به‌عنوان خطاهای اندازه‌گیری برای برامد دو حالتی و متغیرهای کمکی در مدل‌های خطی تعمیم‌یافته، حتی با متغیرهای رده‌ای پنهان، نشان دادند. در واقع، پارامترهای مرتبط با متغیرهای پنهان می‌توانند با استفاده از MI براورد شوند، زیرا مقادیر متغیرهای پنهان همیشه گم شده‌اند و لذا می‌توان آن‌ها را با یک رهیافت ماندگار مانند MI تیمار کرد.

پژوهش‌های آینده باید این مطالعه‌ی مقایسه‌ای را با گنجاندن متغیرهای پیوسته و رسته‌ای هر دو در مدل‌های لوژستیک، با به کارگیری MI و EM برای داده‌های چندمتغیره‌ی گم‌شده تحت حالت‌های دیگری به جز MCAR یا MAR و با بررسی نتایج حاصل از بد مشخص کردن مدل‌های لوژستیک و/یا مدل‌های شامل اثر متقابل بین متغیرهای کمکی، گسترش دهند.

مرجع‌ها

[1] Allison, P.D. (2001). *Missing Data* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-136) Thousand Oaks, CA: Sage.

[2] Buu, Y.P.A. (1999). *Analysis of Longitudinal Data with Missing Values: A Methodological Comparison*. Unpublished doctoral dissertation, Indiana University.

[3] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statisitcal Society*, **39**, 1-38.

[4] Didelez, V. (2002). ML and semiparametric estimation in logistic models with incomplete covariate data. *Statistica Neerlandica*, **56**, 330-345.

[5] Dworkin, R.J. (1987). Hidden bias in the use of archival data. *Evaluation & the Health Professions*, **10**, 173-185.

[6] Ezatti-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., Rubin, D.B. and Schafer, J.L. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. In *Proceedings of the Annual Research Conference* (PP. 257-266). Washington, DC: Bureau of the Census.

[7] Graham, J.W. and Schafer, J.L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed), *Statistical Strategies for Small-Sample Research* (PP. 1-29). Thousand Oaks, CA: Sage.

[8] Hall, J. (1993). *Skills Training for Pregnant and Parenting Adolescents* (NIDA Monograph No. 156, NIH Publication No. 95-3908). Rockville, MD: National Institute on Drug Abuse.

[9] Horton, N.J. and Laird, N.M. (1998). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, **8**, 37-50.

[10] Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, **85**, 765-769.

[11] Kellam, S., Rebok, G., Ialongo, N. and Mayer, L. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry*, **35**, 259-281.

[12] Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, **87**, 1227-1237.

[13] Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[14] Louis T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226-233.

[15] Mason, M.J. (1999). A Review of procedural and statistical methods for handling attrition and missing data in clinical research. *Measurement and Evaluation in Counseling and Development*, **32**, 111-118.

[16] Peng, C.Y.J., Harwell, M., Liou, S.M. and Ehman, L.H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), *Real Data Analysis* (PP. 31-78). Greenwich, CT:Information Age Publishing.

[17] Raghunathan, T.E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, **25**, 99-117.

[18] Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

[19] Rubin, D.B. (1987). *Multiple Imputation for Nonresponses in Surveys*. Wiley, New York.

[20] SAS Institute, Inc. (2004). SAS/STAT® 9.1 User's Guide [Computer manual]. Cary, NC: Author.

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

[22] Schafer, J.L. (1999). Multiple Imputation : A primer. *Statistical Methods in Medical Research*, **8**, 3-15.

[23] Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7**, 147-177.

[24] Schafer. J.L. and Olsen, M.K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research,* **33**, 545-571.

شیرین گلچی
فوق لیسانس آمار
تهران، خیابان فاطمی، خیابان باباطاهر، خیابان سرتیپ فکوری، شماره‌ی ۱۴۵، پژوهشکده‌ی آمار.
رایانشانی: *golchishirin@yahoo.com*

ساناز مهندسی
لیسانس آمار
تهران، خیابان فاطمی، خیابان باباطاهر، خیابان سرتیپ فکوری، شماره‌ی ۱۴۵، پژوهشکده‌ی آمار.
رایانشانی: *s.mohandesi@srtc.ac.ir*