بررسی روشهای برخورد با داده‌های گمشده

افشین آشفته[۱]

چکیده

در این مقاله، روشهای جایگذاری داده‌های گمشده مانند جانهی چندگانه، الگوریتم EM [۲] و داده افزایی[۳] مرور و توضیح داده می‌شوند. شاید تا چندی پیش روش حذف معمول‌ترین روشی بود که برای برخورد با داده‌ها گمشده کاربرد داشت، ولی در حال حاضر، با وجود این روشهای جدید و قدرتمند آماری می‌توان برآوردهای خوبی را برای آنها به دست آورد.

واژه‌های کلیدی: داده گمشده تصادفی، داده گمشده کاملاً تصادفی، زنجیر مارکف، الگوریتم *EM*، داده افزایی، استنباط جانهی چندگانه.

۱. مقدمه

داده گمشده در بیشتر تحقیقات علوم رفتاری و اجتماعی یک شکل آشنا است. در بیشتر مواردی که تحقیق توسط پرسشنامه انجام می‌گیرد، پاسخ دهندگان ممکن است تمایلی به پاسخ دادن نداشته باشند و یا به علل مختلف مانند کمی وقت، بیماری و یا مشکلات دیگر سوالات را بی‌پاسخ بگذارند. در این صورت، در مجموعه داده‌ها، داده گم شده به وجود می‌آید که برای جلوگیری از مشکلاتی مانند کمی حجم نمونه و یا عدم حضور واحدهایی از نمونه در مجموعه داده‌ها باید با استفاده از روشهای مناسب جایگذاری شوند.

۲. چند روش در برخورد با داده گمشده

۱- حذف داده گمشده از تحلیل.

۲- جایگذاری داده گمشده با مقادیر مناسب مانند میانگین.

۳- استفاده از برآوردهای رگرسیونی.

۴- الگوریتم EM [۴] که روشی محاسباتی است برای به دست آوردن برآوردی کارآمد.

۵- روش داده افزایی[۵].

۶- روش جانهی چندگانه[۶] که روشی مبتنی بر میانگین‌های شبیه‌سازی است.

۲-۱- روش حذف داده‌های گمشده

در تجزیه و تحلیلهای چند متغیره با تعداد زیادی متغیر، روش حذف بسیار ناکارآمد است، چراکه تعداد کمی از واحدهای نمونه دارای داده کامل برای تمامی متغیرها می‌باشند. پس، تنها به علت بی‌پاسخ بودن یک متغیر یک واحد نمونه از تحلیل خارج می‌شود. از طرف دیگر، موجب اریب شدن برآوردها می‌گردد، چراکه هیچ تضمینی وجود ندارد که واحدهای باقیمانده نماینده خوبی برای کل داده‌ها باشند.

۲-۲- جایگذاری داده گمشده

جایگذاری داده گمشده با میانگین می‌تواند رابطه بین متغیرها را به طور جدی تحت تاثیر قرار دهد، ولی استفاده از برآوردهای رگرسیونی این

---

[۱] گروه آمار، دانشگاه اصفهان

[۲] Expectation- Maximization Alogorithm

[۳] Multiple Imputation

[۴] The EM Algorithm

[۵] Data augmentation

[۶] Multiple imputations

رابطه را حفظ می‌کند. حتی اگر جایگذاری به طوری انجام گیرد که رابطه بین متغیرها و توزیع احتمال آنها به طور کامل حفظ گردد، باز هم این روش مقادیر دقیقی از داده‌های گمشده را فراهم نمی‌کند، چرا که در این روش مقادیر برآورد شده جایگزین مقادیری ناشناخته می‌شوند که باید حالت احتمالی در برآورد آنها در نظر گرفته شود و نمی‌توان بدون در نظر گرفتن غیر حتمی برآوردهای داده‌های گمشده تقریب خوبی از آنها را جایگذاری کرد.

۲-۳- روشهای جایگذاری چندگانه

- الگوریتم EM
- داده افزایی
- استنباط جانهی چندگانه

۲-۳-۱- الگوریتم EM

در اواخر سال‌های ۱۹۷۰ روبین، دمپستر و لارد[۷] الگوریتم EM را معرفی و گسترش دادند که روشی محاسباتی برای برآورد داده‌های گمشده بود. ثابت شد که این روش یک روش محاسباتی بسیار کارا است و نکته قابل ذکر دیگر تاثیر قابل توجه این روش بر دیدگاه آماردانان بود. تا آن زمان داده‌های گمشده به عنوان عاملی مزاحم بودند که می‌بایست با روشهایی مانند حذف یا جایگذاری از دست آنها خلاص شد، ولی از آن به بعد آماردانان به داده‌های گمشده به عنوان یک منبع تغییرات که نیاز به محاسبه مکرر و روشهای محاسبه احتمالی دارد، نگاه کردند.

در هر مجموعه داده گمشده، مقادیر مشاهده شده اطلاعات غیر مستقیمی را در مورد مقادیر احتمال داده‌های گمشده در اختیار می‌گذارند که اگر با فرضهای اساسی که توضیح داده خواهد شد، ترکیب گردند توزیع احتمال را برای مقادیر گمشده نتیجه می‌دهند که باید در تحلیلهای آماری میانگین گیری شود. روش EM این میانگین گیری را انجام می‌دهد.

انتخاب نام EM به این علت است که در هر تکرار الگوریتم یک مرحله امید ریاضی گیری[۸] و بعد از آن یک ماکسیمم‌سازی[۹] انجام می‌گیرد [۲]. الگوریتم EM از رابطه بین داده‌های گمشده و پارامترهای نامعلوم مدل داده‌ها استفاده می‌کند. بدین صورت که اگر داده‌های گمشده را بدانیم، در این صورت می‌توانیم با استفاده از آنها پارامترهای مدل را مستقیماً برآورد کنیم. همچنین، اگر پارامترهای مدل را داشته باشیم، می‌توان داده‌های گمشده را برآورد کرد. این رابطه بین پارامترهای مدل و داده‌های گمشده ما را بر آن می‌دارد که ابتدا مقادیری را برای پارامترها فرض کرده و مقادیر گمشده را برآورد کنیم. سپس با جایگذاری آنها پارامترها را برآورد کنیم و این کار را تا همگرایی در برآوردها ادامه دهیم. در این صورت برآوردهای به دست آمده با برآوردهای ماکسیمم درستنمایی با در نظر داشتن میانگینی از توزیع‌های مقادیر گمشده یکسان خواهند بود. اگر درصد داده‌های گمشده زیاد باشد، در این صورت همگرایی با تکرار بیشتری حاصل می‌شود.

یک راه برای بررسی همگرایی در EM آزمایش صعودی بودن تابع لگاریتم درستنمایی در هر مرحله است و همچنین بهتر است با مقادیر شروع مختلف و به دست آوردن نتایج یکسان اطمینان حاصل کنیم که لگاریتم درستنمایی یک ماکسیمم منحصر به فرد دارد، زیرا ممکن است دارای مدهای چندگانه باشد و این در صورتی اتفاق می‌افتد که تعداد نمونه کم و یا تعداد داده‌های گمشده زیاد و یا تعداد پارامترهای مدل زیاد باشند.

۲-۳-۲- داده افزایی[۱۰]

مانند روش EM روش داده افزایی یک محاسبه تکراری است که به طور متناوب داده‌های گمشده را جایگذاری کرده و پارامترهای ناشناخته را به صورت یک روند تصادفی پیش‌بینی می‌کند. داده افزایی ابتدا بر اساس مقادیر فرضی پارامترها یک جایگذاری ابتدایی را برای داده‌های گمشده در نظر می‌گیرد و سپس پارامترهای جدید را توسط توزیع پسین به دست آمده از داده‌های کامل برآورد می‌کند. این رویه شبیه‌سازی متناوب پارامترها و داده‌های گمشده یک زنجیر مارکف تولید می‌کند که سرانجام تثبیت شده یا در توزیع همگرا می‌شود.

---

[۷] Rubin, Dempster, Laird

[۸] Expectation

[۹] Maximazation

[۱۰] Data Augmentation

توزیع پارامترها به یک توزیع پسین که در اصل میانگینی از داده‌های گمشده در آن نقش دارند، می‌رسد. همچنین، توزیع داده‌های گمشده نیز به سمت توزیع پیش‌بینی کننده‌ای که در حقیقت همان چیزی است که برای به دست آوردن جانهی‌های چندگانه دقیق لازم است، پیش می‌رود. دقیقاً مانند EM در داده افزایی نیز تعداد تکرار لازم برای همگرایی به میزان اطلاعات داده‌های گمشده ربط دارد.

توجه شود که همگرایی در داده افزایی با همگرایی در روش EM متفاوت است. چرا که در EM همگرایی زمانی حاصل می‌شود که پارامترها از یک تکرار تا تکرار بعدی تغییر نکنند، ولی همگرایی در داده افزایی زمانی است که توزیع پارامترها از مرحله‌ای به بعد در صورت تغییر مقادیر تصادفی پارامتر ثابت باقی بماند. به همین علت، تشخیص همگرایی در داده افزایی پیچیده‌تر از روش EM است.

- تشخیص همگرایی در داده افزایی

همگرایی در داده افزایی به صورت تاخیری در همبستگی دنباله‌ای تعریف می‌شود. به این صورت که داده افزایی در تکرار k ام همگرا خواهد شد. اگر مقدار هر پارامتر در تکرار t ام به صورت آماری مستقل از مقدار آن در تکرار (t+k) ام برای $t = ۱,۲,۳,...$ باشد.

ابتدا مقادیر پارامترها را در هر تکرار ثبت کرده، سپس با نمودار تابع خود همبستگی[۱۱] سری زمانی می‌توان به راحتی ضریب همبستگی پیرسن تاخیر k ام را برای مقادیر مختلف k به دست آورد، یعنی همبستگی بین مقادیر شبیه‌سازی یک پارامتر در هر تکرار و مقادیر آن در k تکرار قبل. اگر مقادیر تابع خود همبستگی بعد از تاخیر k برای تمام پارامترها از k ام به صفر برسند، در این صورت می‌گوییم داده افزایی در تکرار k ام همگرا است. نکته جالب اینجاست که روش داده افزایی سریعتر از EM به همگرایی می‌رسد.

۲-۳-۳- استنباط در مورد جانهی چندگانه

روبین در سال ۱۹۸۷ نمونه‌ای از جانهی چندگانه را گسترش داد که بر میانگین گیری بر اساس شبیه‌سازی استوار بود.

در روش جانهی چندگانه هر مقدار گمشده با مجموعه‌ای از مقادیری که از توزیع پیش‌بینی شده، به دست آمده‌اند، جایگذاری می‌شوند.

تغییرات موجود بین m جایگذاری، نشان دهنده این است که فرض عدم وجود حتمیت برای مقادیر پیش‌بینی در نظر گرفته شده است. بعد از به کار بردن جانهی چندگانه به تعداد m مجموعه داده‌های کامل خواهیم رسید که هر کدام را می‌توان توسط روشهای آماری موجود تحلیل کرد و نتایج به دست آمده از هر مجموعه داده را توسط قوانین ساده ارائه شده روبین در سال ۱۹۸۷ ترکیب کرد.

روش جانهی چندگانه به چند دلیل جالب است؛

- □ این روش برای به کار بردن تحلیلهای آماری و نرم‌افزارهایی که نیاز به مجموعه داده کامل دارند، بسیار مناسب است.
- □ هر کدام از m جایگذاری می‌توان در تحلیل‌های مختلف استفاده شود و نیازی به تولید دوباره مجموعه‌های داده‌ها توسط روش جانهی چندگانه برای انجام تحلیل‌های جدید وجود ندارد.
- □ انحراف معیارها و مقادیر احتمال[۱۲] و غیره که از روش جانهی چندگانه به دست می‌آیند، معتبر خواهند بود. چرا که عدم وجود حتمیت برای داده‌های جایگذاری شده را در نظر گرفته‌ایم.
- □ این روش در تعداد کمی از تکرار کارآیی بسیار بالایی برای به دست آوردن نتایج دقیق ارائه می‌دهد (در اغلب موارد فقط ۳ تا ۵ تکرار کافی است!).

شاید این تعداد تکرار کم در تولید مجموعه‌های داده‌ها باعث تعجب شود، ولی روبین [۸، صفحه ۱۱۴] نشان داده است که کفایت یک برآورد بر اساس m جایگذاری تقریباً برابر است با

$$\left(۱+\frac{\gamma}{m}\right)^{-۱}$$

که در آن، γ کسری از اطلاعات گمشده[۱۳] با توجه به مقادیر برآورد شده است. یعنی، در صورت عدم وجود داده گمشده چه مقدار ممکن

[۱۱] ACF

[۱۲] P-Value

[۱۳] Fraction of Missing Information

است (با توجه به برآورد موجود) دقت برآورد بیشتر شود. روش محاسبه γ در انتها نشان داده خواهد شد. کارآیی به دست آمده از مقادیر مختلف γ و m در جدول ۱ نشان داده شده است و می‌توان دید که کارآیی بالایی بعد از تعداد کمی تکرار به دست می‌آید.

روش جانهی چندگانه در حدود ۲۰ سال قبل [۷] برای اولین بار ارائه شد، ولی در طی این سالها این روش به علت عدم وجود نرم‌افزار مناسب برای تولید جانهی‌های چندگانه چندان مورد استفاده قرار نگرفت. چرا که به دست آوردن توزیع احتمال، که داده‌ها باید از آن تولید شوند، پیچیده و بغرنج به نظر می‌رسید. اخیراً با معرفی روشهای شبیه‌سازی که به نام زنجیر مارکف مونت‌کارلو[14] معروف هستند، تحولی در مدلهای پارامتری کاربردی ایجاد شده است [٤].

حال به بررسی دقیق‌تر روشهای جایگذاری و فرضیات لازم می‌پردازیم.

۳. فرضیات

برای بررسی فرضیات لازم است تا اطلاعات کاملی نسبت به موارد زیر داشته باشیم.

- مدل داده‌ها
- توزیع پسین برای پارامترهای مدل
- چگونگی و خصوصیات مقادیر بدون پاسخ.

۳-۱- مدل داده‌ها

تحلیل‌گران می‌دانند که احتمال برازنده شدن مدلهای ساده‌ای مانند نرمال چند متغیره به داده‌ها بسیار کم است. ولی در روشهای جانهی چندگانه نشان داده شده است که حتی با این فرض تقریبهای بسیار خوبی به دست می‌آید.

برای مثال، فرض کنید بر روی متغیرهایی ترتیبی یا دو حالتی کار می‌کنیم. اغلب قابل قبول است که مقادیر را با فرض نرمال برآورد و سپس به نزدیکترین گروه یا عدد گرد کنیم. متغیرهای با توزیع شدیداً چوله را می‌توان توسط تبدیلاتی مانند لگاریتم به نرمال نزدیک کرد.

مطالعات شبیه‌سازی برای روشهای استوار[15] جانهی چندگانه توسط ازتی و رایس[16] [۳] و شیفر[17] [۱۰] انجام گرفته است.

دو نکته را باید مورد توجه قرار داد

۱. مدل در نظر گرفته شده باید با تحلیلی که می‌خواهیم بر مجموعه‌های داده‌ها انجام دهیم، سازگار باشد (حداقل به طور تقریبی).

۲. مدل در نظر گرفته شده باید روابط متغیرها را حفظ کند.

برای توضیح بیشتر مثال ساده زیر را در نظر می‌گیریم.

مثال: فرض کنید Y با یک مدل نرمال و با در نظر گرفتن متغیر X_1 جایگذاری شده است. بعد از جایگذاری می‌خواهیم از رگرسیون خطی برای پیش‌بینی Y توسط X_1 و متغیر دیگری مانند X_2 که در مدل جایگذاری حضور نداشته است، استفاده کنیم. در این صورت، برآورد ضریب رگرسیونی X_2 به سمت صفر میل می‌کند. چرا که Y تنها بر اساس متغیر اول جایگذاری و پیش‌بینی شده است. پس لازم است که در مدل جایگذاری تمامی متغیرهای مورد نیاز در تحلیل حضور داشته باشند.

۳-۲- توزیع پسین

تئوری جانهی چندگانه قوانین اساسی از احتمال که به نظریه بیز مشهور است را شامل می‌شود. برای این منظور نیاز به توزیع پسین برای پارامترهای وارد شده در مدل داریم که در صورت بزرگ بودن حجم نمونه تقریباً تمامی توزیع‌های توجیه‌پذیر پسین منجر به نتایج یکسان می‌شوند. در غیر این صورت همانطور که می‌دانیم هر توزیع پسینی نتایج متفاوتی خواهد داد که باید به طور دقیق تشخیص داده شود (برای مطالعه روش انتخاب توزیع پسین می‌توان به مرجع [۱۰] مراجعه کرد).

۳-۳- مکانیزم مقادیر بی پاسخ

در روشهای جایگذاری فرض می‌شود که سیستم از دست رفتن داده‌ها قابل چشم‌پوشی[18] [۸] و یا دارای فرضیات گمشدگی تصادفی[19] است [٦].

[14] Makov chain Monte Carlo (MCMC)

[15] Robust

[16] Ezzati and Rice

[17] Schafer

[18] Ignorable

[19] Missing at Random (MAR)

ممکن است در نگاه اول از گمشدگی تصادفی این گونه استنباط شود که مقادیر گمشده باید یک ریز نمونه تصادفی از مجموعه کل داده‌ها باشند. در صورتی که این تعریف مربوط به داده‌های گمشده کاملاً تصادفی است و اغلب در دنیای واقعی اینگونه نیست. داده‌های گمشده را کاملاً تصادفی گوییم، اگر احتمالات متغیر پاسخ، به داده‌های گمشده یا مشاهده شده ربطی نداشته باشد.

گمشدگی تصادفی به این معنی است که احتمالات از دست رفتن داده‌ها با مقادیر داده‌هایی که مشاهده شده‌اند، ربط داشته ولی با مقادیر گمشدگی رابطه‌ای نداشته باشند. برای روشن‌تر شدن موضوع مجموعه داده‌های دو متغیره را در نظر بگیرید که X همیشه مشاهده می‌شود و Y ممکن است گاهی دارای داده گمشده باشد. تحت فرض گمشدگی تصادفی می‌توان Y را به وسیله X و روش رگرسیون خطی پیش‌بینی کرد و داده‌های گمشده را با استفاده از این مدل رگرسیون جایگذاری نمود. در واقع، گمشدگی تصادفی کمک می‌کند تا احتمالات از دست رفتن مشاهدات Y را تنها با مقادیر داده‌هایی از X که مشاهده شده‌اند، مرتبط دانسته و مطمئن باشیم که مقادیر گمشده تاثیری در احتمالات نمی‌گذارند.

در حقیقت، گمشدگی تصادفی فرضیه‌ای است که به ما اجازه می‌دهد تا ابتدا برآوردی از ارتباط بین متغیرها برای داده‌های مشاهده شده به دست آورده و سپس این ارتباط را برای به دست آوردن برآوردهایی نا اریب از مقادیر گمشده به کار بریم (این مطلب طی چند قضیه با فرض شرایط گمشدگی تصادفی در مرجع [٦] آمده است).

نمادهای زیر را که در بیشتر مقالات مورد استفاده قرار گرفته‌اند، در نظر بگیرید؛

$U=(U_1,U_2,\dots,U_n)$ یک بردار متغیرهای تصادفی با تابع چگالی احتمال f_θ باشد. هدف به دست آوردن استنباطی در مورد θ (بردار پارامترهای تابع چگالی احتمال f_θ است). $M=(M_1,M_2,\dots,M_n)$ بردار متغیرهای تصادفی نشانگر داده گمشده است که در آن M_i دو مقدار صفر و یک را اختیار می‌کند. احتمال اینکه M برابر با $m=(m_1,m_2,\dots,m_n)$ باشد، به شرط اینکه U مقدار $u=(u_1,u_2,\dots,u_n)$ را اختیار کند، برابر با $g_\phi(m|u)$ است که در آن ϕ بردار پارامترهای مزاحم توزیع است. توزیع شرطی $g_\phi(m|u)$ دلالت بر جریان به وجود آورنده داده گمشده[20] دارد. اگر $m_i=1$ مقدار متغیر تصادفی U_i مشاهده شده و اگر $m_i=0$ مقدار آن گمشده است. بردار U را به دو قسمت $U_{(1)}$ و $U_{(0)}$ تقسیم می‌کنیم که در آن $U_{(0)}$ بردار مقادیر گمشده و $U_{(1)}$ بردار مقادیر مشاهده شده می‌باشد. یعنی، $U=(U_{(0)},U_{(1)})$.

در این صورت داده‌های گمشده دارای خاصیت گمشدگی تصادفی هستند. اگر برای هر مقدار ϕ، $g_\phi(\tilde{m}|\tilde{u})$ مقدار یکسانی برای تمام $u_{(0)}$ اختیار کند. یعنی برای هر مقدار ممکن پارامتر ϕ، احتمال شرطی الگوی داده گمشده به شرط داده‌های گمشده، به علاوه مقادیر مشاهده شده برای تمام مقادیر داده‌های گمشده یکسان باشد.

داده‌ها دارای خاصیت مشاهده شدگی تصادفی[21] هستند، اگر برای هر مقدار ϕ و $u_{(0)}$، $g_\phi(\tilde{m}|\tilde{u})$ مقدار یکسانی را برای تمامی $U_{(1)}$ اختیار کند.

حال فرض کنید که

۱. داده‌های گمشده دارای خاصیت گمشدگی تصادفی هستند

۲. داده‌ها دارای خاصیت مشاهده شدگی تصادفی هستند.

در این صورت با فرض $k_{\theta,\phi}(\tilde{m})>0$ داریم،

$$\int f_\theta(u)du_{(0)}=\int\left\{\frac{f_\theta(u)g_\phi(\tilde{m}|u)}{k_{\theta,\phi}(\tilde{m})}\right\}du_{(0)}$$

که در آن، $k_{\theta,\phi}(\tilde{m})=\int f_\theta(u)g_\phi(\tilde{m}|u)du$ می‌باشد. یعنی احتمال حاشیه‌ای $M=\tilde{m}$.

سمت راست تساوی نشان دهنده توزیع نمونه مورد بررسی ما با در نظر گرفتن جریان به وجود آورنده داده گمشده است و سمت چپ تساوی نشان دهنده توزیع نمونه‌ای با توجه به داده‌های مشاهده شده می‌باشد. دیده می‌شود که با دو شرط (۱) و (۲) می‌توان این دو توزیع را برابر در نظر گرفت. چرا که با توجه به این دو شرط $g_\phi(\tilde{m}|u)$ برای هر ϕ مقدار یکسانی را برای تمامی uها اختیار می‌کند (توجه کنید که این مطلب استقلال توزیع‌های U و M را فقط در صورتی که موضوع برای تمامی $\tilde{m}$ برقرار باشد، نتیجه می‌دهد). بنابراین،

---

[20] Process that causes missing data

[21] Observed at Random (OAR)

$$k_{\theta,\phi}(\widetilde{m}) = \int f_\theta(\widetilde{m}|u)du = g_\phi(\widetilde{m}|u)\int f_\theta(u)du = g_\phi(\widetilde{m}|u) = c$$

$$\int\left\{\frac{f_\theta(u)g_\phi(\widetilde{m}|u)}{k_{\theta,\phi}(\widetilde{m})}\right\}du_{(0)} = \int\left\{\frac{f_\theta(u)g_\phi(\widetilde{m}|u)}{g_\phi(\widetilde{m}|u)}\right\}du_{(0)} = \int f_\theta(u)du_{(0)}$$

بر اساس فرضیه گمشدگی تصادفی، چند نکته را باید در نظر داشت.

- فرض گمشدگی تصادفی بر اساس چگونگی روابط متغیرهای موجود در مجموعه‌های داده‌ها قابل قبول است. بدین معنی که اگر متغیر X با داده‌های گمشده متغیرهای دیگر و همچنین مقادیر مشاهده شده آنها ارتباط داشته باشد و در مجموعه‌های داده‌ها موجود نباشد، فرض گمشدگی تصادفی برای استفاده از روش جایگذاری درست نیست. به همین دلیل در یک روند جایگذاری تمامی متغیرهای مرتبط با مقادیر گمشده را می‌بایست در نظر گرفت.
- فرض گمشدگی تصادفی را نمی‌توان با داده‌های موجود آزمون کرد، مگر اینکه داده‌های گمشده را نیز داشته باشیم. بنابراین، نمی‌توان برقرار بودن فرض گمشدگی تصادفی را به راحتی بررسی کرد، ولی به علت اینکه بتوان از روشهایی آماری برای پیش‌بینی مقادیر گمشده بر اساس داده‌های موجود استفاده کرد، در بیشتر مواقع این فرض را برقرار در نظر می‌گیریم.

در حال حاضر، روشی مبنی بر عدم نادیده گرفتن داده‌های گمشده وجود ندارد و تا زمانی که روشی پیشنهاد گردد، بهتر است از فرضهایی مانند گمشدگی تصادفی که داده‌های گمشده را نادیده گرفته و پیش‌بینی را بر اساس داده‌های موجود انجام می‌دهد، استفاده گردد. در ضمن اینکه گمشدگی تصادفی برآوردهای بسیار بهتری نسبت به روشهایی مانند حذف یا جایگذاری میانگین پیشنهاد می‌کند (برای توضیحات بیشتر در مورد روشهای عدم نادیده گرفتن داده‌های گمشده می‌توان به مرجع [۱۰] صفحات ۲۷ تا ۲۸ مراجعه کرد).

۴. قوانین جانهی چندگانه

با استفاده از قوانین جانهی چندگانه می‌توان تمامی تحلیلهای آماری که نیاز به مجموعه داده کامل دارند را انجام داد. فرض کنیم رگرسیون خطی یا لوژستیک را توسط نرم‌افزارهای آماری انجام می‌دهیم. هر مدل باید m بار یعنی به تعداد جایگذاریهای داده‌های گمشده برازش شود. در این صورت، نتایج به دست آمده از مجموعه‌های کامل داده‌ها، انعکاسی از عدم وجود حتمیت برای داده‌های گمشده را دربرخواهد داشت. حال، برای به دست آوردن ضرایب و انحراف معیارهای برآورد شده از قوانین روبین برای ترکیب نتایج و به دست آوردن برآوردهای نهایی استفاده می‌شود [۸].

قوانین روبین به صورت زیر است؛

فرض کنید $\hat{Q}$ برآورد کیفیت مورد علاقه جامعه و U برآورد واریانس آن باشد. برای مثال، می‌توان $\hat{Q}$ را برآورد ضریب رگرسیونی و U را واریانس آن در نظر گرفت.

بعد از انجام تحلیل یکسان بر روی هر مجموعه داده کامل شده به تعداد m برآورد $\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_m$ و همچنین واریانسهای آنها $U_1,U_2,\ldots,U_m$ خواهیم داشت. برآورد جانهی چندگانه به صورت $\overline{Q}=\frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$ است.

واریانس کل برآورد دارای دو مؤلفه است که یکی تغییرات درونی هر مجموعه داده‌ها (واریانس درون- جایگذاری) و دیگری تغییرات بین مجموعه‌های داده (واریانس بین - جایگذاری) را مشخص می‌کند.

واریانس درون- جایگذاری برابر است با، $\overline{U}=\frac{1}{m}\sum_{i=1}^{m}U_i$ که میانگین واریانسهای برآورد شده است.

واریانس بین- جایگذاری به صورت $B=\sum_{i=1}^{m}\left(\hat{Q}-\overline{Q}\right)^2/(m-1)$ است که واریانس نمونه‌ای خود برآوردها می‌باشد. واریانس کل (T) مجموع این دو مولفه به علاوه ضریب تصحیحی است که برای خطای شبیه‌سازی $\overline{Q}$ در نظر گرفته می‌شود.

$$T=\overline{U}+\left(1+\frac{1}{m}\right)B$$

توجه کنید که اگر داده گمشده نداشته باشیم $\hat{Q}_1,\hat{Q}_2,\ldots,\hat{Q}_m$ مشخص و برابرند و $B=0$ و T برابر با $\overline{U}$ خواهد بود که $\overline{U}$ با هر کدام از U_iها برابر است.

فاصله اطمینان ۹۵ درصد به صورت $\overline{Q} \pm 2\sqrt{T}$ به دست می‌آید که به صورت دقیق‌تر برابر است با

$$\overline{Q} \pm t_{df}\sqrt{T} \qquad , \qquad df = (m-1)\left(1+\frac{m\overline{U}}{(m+1)B}\right)^{2}.$$

توجه کنید که اگر تعداد تکرارها بی‌شمار باشد $(m=\infty)$ واریانس کل بـه مجمـوع دو مولفـه واریـانس کـاهش می‌یـابد و فاصلـه اطمینان بر اساس توزیع نرمال خواهد بود. اگر مقدار درجه آزادی خیلی کوچک باشد (کمتـر از ۱۰) بـا افزایـش تعـداد تکرارهـا می‌تـوان دقـت برآوردها را افزایش و طول فاصله اطمینان را کاهش داد.

روبین نشان داده است که یک برآورد برای نسبت اطلاعات گمشده برای کمیت جامه Q برابر است با

$$\gamma = \frac{r+\dfrac{2}{df+3}}{r+1}$$

و در آن $r = \left(1+m^{-1}\right)B/\overline{U}$ که مشخص می‌کند چه مقدار برآورد Q ممکن است توسط داده‌های گمشده تحت تاثیر قرار گیرد [۸].

۵. نکاتی در مورد روش جانهی چندگانه

روش جانهی چندگانه یکی از روشهای برخورد با داده‌های گمشـده است و لزوما در تمامی شرایط بهترین روش نیست. برای مثال، زمانی که مدل‌های کاملا پارامتری در اختیار داریم، اغلـب برآوردهـای ماکسیمم درستنمایی را می‌توان به راحتی با استفاده از روشهای عددی مـانند EM محاسبه کرد.

همان طور که می‌دانیم روش حذف داده‌های گمشده بسـیار ساده‌تر از روش جانهی چندگانه است. اگر داده‌های مشاهده شده نماینده خوبی از کل داده‌ها باشد و سهم داده‌های گمشده از کـل داده‌هـا زیـاد نباشـد، شاید روش حذف مناسـب بـه نظر برسـد. لیکـن، در کـل روش حـذف داده‌های گمشده تنـها زمـانی کـارآیی لازم را دارد کـه دارای داده‌هـای گمشده کاملا تصادفی باشیم. به بیان دیگر، روش حذف فـرض می‌کنـد که واحدهای حذف شده یکـی زیـر نمونـه تصـادفی هسـتند. وقتـی کـه واحدهای حذف شده به طور سیستماتیک با بقیه داده‌ها متفاوت باشـند، با روش حذف برآوردها شدیدا اریب خواهند شد. بـه عـلاوه، همـانطور که گفته شد در تحلیلهای چند متغیره روش حذف شدیدا حجم نمونه را کاهش می‌دهد.

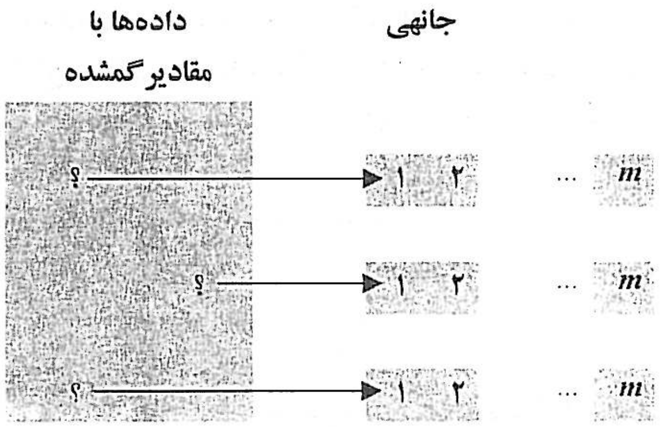
در حقیقت روش جانهی چندگانه شباهت نزدیکی با الگوریتـم EM و روشهای دیگر عددی دارد که برآوردهای ماکسیمم درستنمایی را بـر اسـاس تنـها مقـادیر مشـاهده شـده محاسـبه می‌کننـد. ایـن روشـها تـابع درستنمایی که در اصل متوسطی است بر توزیع‌های پیش‌بینی شده، برای داده‌های گمشده را محاسبه می‌کنند. روش جانهی چندگانه شـبیه همیـن متوسط‌گیری را به جای روشـهای عـددی بـه وسیله مونت‌کارلو انجام می‌دهد.

شاید این سوال مطرح شود که آیا جانهی چندگانه ربطـی بـه زنجیر مارکف مونت‌کارلو دارد یا خیر؟ در جواب این سوال باید گفت، زنجیر مارکف مونت‌کـارلو یـک روش شبیه‌سـازی اسـت کـه بـرای بـه دسـت آوردن مقـادیر تصـادفی از توزیع‌هـای غـیر اسـتاندارد بـه وسـیله زنجـیر مارکف به کار می‌رود. این روش با شبیه‌سازی پارامترهـا از یـک توزیـع پسین بیزی برای یک مدل پارامتری پیچیده در اصل این امکان را بـه مـا می‌دهـد تـا در روش جانـهی چنـدگانـه تعـدادی مقـادیر مسـتقل بـرای داده‌های گمشده توسط توزیعی که برآورد شده است، به دست آوریـم و از این مقادیر برای استنباط جانهی چندگانه استفاده کنیم.

توجه کنید که فرض گمشدگی تصادفی را به صورت ساده می‌توان اینگونه تصور کـرد کـه ایـن فـرض بـه مـا ایـن اطمینـان را می‌دهـد کـه داده‌های گمشده دارای هیچ اطلاعاتی از توزیع احتمال مقـادیر گمشـده نبوده‌اند. می‌توان گفت که اگر بی‌پاسخ ماندن متغیری به داده‌های فعلـی یا قبلی آن ربط نداشته باشد، فرض گمشدگی تصادفی قابل دفاع است. در آخر می‌توان گفت که روش جانهی چندگانه یـک روش سـاختن از هیچ چیز است. در اصل، این روش نیز مانند روشـهای دیـگر شبیه‌سـازی تنها فرض احتمالی داده‌های جایگزین را در نظر گرفته و نمی‌توان گفت که معجزه‌ای بـرای دیـدن داده‌هـایی کـه در اصـل گمشـده‌اند، رخ داده است. البته نباید فراموش کـرد کـه در نظر گرفتـن همیـن فـرض نـیز در محاسبات تأثیر به سزایی دارد و بسیار مهم است.

جدول ۱

			γ		
m	۰/۱	۰/۳	۰/۵	۰/۷	۰/۹
۳	۹۷	۹۱	۸٦	۸۱	۷۷
٥	۹۸	۹٤	۹۱	۸۸	۸٥
۱۰	۹۹	۹۷	۹٥	۹۳	۹۲
۲۰	۱۰۰	۹۹	۹۸	۹۷	۹٦



ماتریس داده‌های چند متغیره با مقادیر گمشده و

جانهی چند گانه

مراجع

[1] Ashofteh, A., Jahanshahi, M.A. and Bozorgnia, A., 2002. *Estimation of Systematic Missing Values by Simulation*, Submitted.
[2] Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. *Maximum Likelihood from Incomplete Data Via the EM Algorithm,* Journal of the Royal Statistical Society, Series B, 39, pp.1-38.
[3] Ezzati-Rice, TM., 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. 25 7-266. Bureau of the Census, Washington, D.C.
[4] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., (Eds.), 1996. *Markov Chain Monte Carlo in Practice,* Chapman & Hall, London.
[5] Meng, X.L. and Rubin, D.B., 1993. *Maximum Likelihood Estimation Via the ECM Algorithm. A General Framework*, Biometrika, 80, pp.267-278.
[6] Rubin, D.B., 1976. *Inference and Missing Data,* Biometrika, 63, pp.581-592.
[7] Rubin, D.B., 1978. *Multiple Imputations in Sample Surveys- A Phenomenological Bayesian Approach to Nonresponse,* Proceedings of the Survey Research Methods Section, American Statistical Association, pp.20-34.
[8] Rubin, D.B, 1987. *Multiple Imputations for Nonresponse in Surveys*, John Wiley & Sons, New York.
[9] Rubin, D.B, 1996. *Multiple Imputations After 18+ Years (with discussion),* J. A. Stat. Asso., 91, pp. 473-489.
[10] Schafer, J.I., 1997. *Analysis of Incomplete Multivariate Data,* Chapman & Hall, London.
[11] Tanner, M.A. and Wong, W.H., 1987. *The Calculation of Posterior Distributions by Data Augmentation (with discussion),* J. Amer. Statist. Asso., 82, pp. 528-550.