نامهٔ آموزش عالی، دورهٔ جدید
سال نهم، شمارهٔ سی و سوم، بهار ۱۳۹۵، صص ۳۷ - ۱۱

Higher Education Letter
Vol. 9, No. 33, Spring 2016

روش‌های برخورد با داده‌های گمشده: مزایا، معایب، رویکردهای نظری و معرفی نرم‌افزارها

Methods of Dealing with Missing Data: Advantages, Disadvantages, Theoretical Approaches and Application of Software

تاریخ دریافت مقاله:۹٤/۱۱/۸

تاریخ پذیرش مقاله:۹۵/۲/۳۰

رقیه باقی‌یزدل *
احسان جمالی **
ابراهیم خدایی ***
مجتبی حبیبی ****

Roghayeh Baghi Yazdel
Ehsan Jamali
Ebrahim Khodaei
Mojtaba Habibi

چکیده: در تحلیل داده‌ها، گاهی برخی مشاهدات به دلایل گوناگون و روش‌های متفاوت، گمشده محسوب می‌شوند. چگونگی برخورد با این مشاهدات در تحلیل داده‌ها، به دلیل اهمیت نتایج حاصل از آنها به‌ویژه در تصمیم‌گیری‌های حساس، از اهمیت به سزایی برخوردار است. پیش از این، برای غلبه بر مشکل داده‌های گمشده مرسوم‌ترین روش، حذف داده‌های گمشده بود که منجر به داده‌هایی با کیفیت پایین و به تبع آن تحلیل و استخراج نتایج دارای سوگیری می‌شد. امروزه با پیشرفت‌های علمی در حوزه‌های گوناگون و پیدایش روش‌های توانمند آماری، می‌توان پیش از مدل‌سازی داده‌های ناکامل، مقادیر گمشده را با مقادیر مناسب جایگذاری یا برآورد کرد. در این مقاله، به بررسی انواع داده‌های گمشده، روش‌های جانهی، مفروضه‌ها، مقایسه روش های جانهی و مزایا-معایب آنها و معرفی مختصر نرم-افزارهای کاربردی در این حوزه پرداخته شده است. برای تحلیل داده‌ها (با استفاده از نرم‌افزار R) یک نمونه داده تجربی مربوط به نتایج نود و یکمین آزمون تولیمو در سال ۱۳۹۳ ارائه شده است. نتایج نشان داد که در خصوص این داده‌ها از بین سه روش جانهی چندگانه، الگوریتم EM و الگوریتم DA، با توجه به معیار MSE، الگوریتم EM، بهترین عملکرد را داشته است.

واژگان کلیدی: داده‌های گمشده، روش‌های جانهی، نرم‌افزارهای کاربردی

Abstract: In some cases, in data analysis, missingness happens in the observation for different reasons and ways. How to deal with these observations in the data analysis process is very important, especially in the high stack decisions, the usual way to dealing with missing data issues was eliminating missing values. This method leads to low quality in data and consequently leads to bias in results. Today, regarding to the advances in various fields of science and powerful statistical methods, imputation methods are applied is possible in the case of incomplete data. In this paper, the various types of missing data, imputation methods, the assumptions, their advantages and disadvantages were discussed. In this setting, we shall try to provide applied examples using statistical software. Especially an applied example due to 91th TOLIMO test, which was provided by Iranian National Organization of Educational Testing, has been presented (NOET). Comparison of results according to MSE using three methods of multiple imputation, the EM algorithm and the DA algorithm, has showed that the EM algorithm had the best performance for analyzing this data set.

Keywords: Missing Data, Imputation Methods, Applications Software's

* کارشناس ارشد امور پژوهشی سازمان سنجش آموزش کشور (نویسنده مسئول baghi_y@yahoo.com)
** استادیار سازمان سنجش آموزش کشور
*** دانشیار دانشگاه تهران
**** استادیار پژوهشکده خانواده دانشگاه شهید بهشتی

مقدمه

در بسیاری از بررسی‌ها و پرسشنامه‌ها با داده‌های گمشده[1] مواجه هستیم. در هر پرسشنامه‌ای تمایل نداشتن به پاسخگویی یا کمبود وقت می‌تواند به ایجاد داده گمشده منجر شود که تهدید عمده‌ای برای صحت نتایج حاصل از تحلیل داده‌ها محسوب می‌شود. بنابراین جانهی[2] مقادیر مناسب به داده‌های گمشده ازجمله چالش های مهم تحلیل آماری است. در نظر نگرفتن داده گمشده یا استفاده از روش‌های ابتدایی یا نامناسب برخورد با این داده‌ها (به‌عنوان مثال جای‌گذاری داده‌های گمشده با میانگین داده‌ها) به مشکلاتی مانند کم شدن حجم نمونه، ایجاد اریبی[3]، حذف اطلاعات ارزشمند (وایمن[4]، ۲۰۰۳)، کاهش کیفیت و سوگیری نتایج منجر می‌شود.

داده گمشده، عنوانی کلی برای مجموعه‌ای از حالت‌ها است و در پرسشنامه می می‌تواند از نوع پاسخ داده نشده[5]، حذف شده[6] و یا سانسور شده[7] باشد. در موارد پاسخ داده نشده، فرد به گزینه مورد سؤال پاسخ نمی‌دهد. مثلاً در خیلی از موارد گزینه درآمد از طرف اشخاص پاسخ داده نمی‌شود. در موارد حذف شده، پژوهشگر پاسخ داده شده به پرسش را بنا به دلایلی حذف می‌کند و یا با آنها مخالفت می‌کند، مثلاً در پاسخنامه کنکور سراسری پاسخ‌های مخدوش توسط دستگاه حذف می‌شود و در موارد سانسور شده پاسخ پرسش‌ها از جایی به بعد وجود ندارد، مثلاً در مواردی که در یک تحقیق داده‌ها از نوع پنلی[8] می‌شوند یکی از گروه‌های پاسخ‌دهنده به پرسشنامه، دیگر به پرسش‌ها پاسخی نمی‌دهد (لیتل و روبین[9]، ۱۹۸۷). تحقیقات پیشین نشان داده است که داده‌های گمشده می‌توانند تأثیرات بدی (شامل برآوردهای اریب برای پارامترها و خطاهای استاندارد متورم[10]) بر تحلیل‌های آماری مبتنی بر پاسخ‌ها داشته باشند (دلیو و همکاران[11]، ۲۰۰۳). کارهای زیادی در زمینه روش‌های جانهی برای داده‌های گمشده انجام شده است (ون بورن[12]، ۲۰۱۲)؛ اما بیشتر این

---

1. Missing data
2. imputation
3. biases
4. Wayman
5. Not responded
6. Omitted
7. Not reach
8. panel data
9. Little & Rubin
10. Inflated standard errors
11. De Leeuw
12. Van Buuren

روش‌ها برای داده‌های پیوسته و غالباً مبتنی بر فرض نرمال بودن پاسخ‌ها است. در پژوهش‌های علوم رفتاری، بهداشت، سلامت، پزشکی و آزمون‌های مبنی بر پرسشنامه‌های زمینه‌یابی، نگرش‌سنج و استعداد سنجی معمولاً با داده‌های پاسخنامه‌ای مواجه هستیم که پاسخ‌ها به صورت رسته‌ای در نظر گرفته می‌شوند.

در این پژوهش، آزمون تولیمو[1] برای بررسی و جانهی داده‌های گمشده در نظر گرفته شده است. این آزمون توسط سازمان سنجش آموزش کشور به‌عنوان آزمونی استاندارد برای تعیین سطح دانش زبان انگلیسی داوطلبان، طراحی و اجرا می‌شود که بیشتر متقاضیان آن را داوطلبان دکتری تشکیل می‌دهند.

در پاسخنامه‌های مربوط به آزمون‌های تولیمو با مواردی از داده گمشده روبرو هستیم. در این آزمون، هر سؤال چهار گزینه دارد که در موارد بی‌پاسخی، داوطلب گزینه پنجم، یعنی نمی‌دانم را انتخاب می‌کند. ازجمله اصلی‌ترین موارد رخداد آن، بی‌پاسخ گذاشتن سؤال از طرف داوطلب است که به دلیل دشواری و یا زمان‌بر بودن سؤال رخ داده است و یا اینکه پاسخ مورد نظر با پاسخ سؤال دیگر ناسازگار بوده و به‌ناچار حذف شده است که این مورد می‌تواند یکی از اشکالات در طرح سؤال باشد. با توجه به اینکه در این آزمون‌ها پاسخ‌های اشتباه دارای نمره منفی نیست، لذا می‌توان مقادیر گمشده را به‌عنوان داده گمشده حذف شده در نظر گرفت و آنها را با مقادیر مناسبی جایگزین کرد. همچنین در مواردی نیز ممکن است داده گمشده به دلیل اشکال در دستگاه ثبت پرسشنامه رخ داده باشد و یا اینکه بی‌پاسخی به دلیل حذف پاسخ مخدوش توسط دستگاه رخ داده باشد، در این موارد نیز که داده گمشده از نوع حذف شده است، می‌توان به جای حذف رکورد، به برآورد داده آنها اقدام کرد.

در مقاله حاضر ابتدا مکانیسم گمشدگی داده‌ها و روش‌های جانهی، مفروضه‌ها و شرایط هر کدام، بررسی و نرم‌افزارهای کاربردی در این حوزه به‌طور مختصر در خصوص برخی از روش‌های جانهی معرفی شده است. سپس با استفاده از یک مثال کاربردی به بررسی داده گمشده در پاسخنامه‌های نود و یکمین آزمون تولیمو در سال ۱۳۹۳ پرداخته شده است و با استفاده از روش‌های معرفی شده به برآورد پاسخ‌های گمشده در این آزمون اقدام شده است. در نهایت مقادیر برآورد شده با مقادیر اصلی مقایسه شده است. این روش اگرچه پاسخگویی به داوطلبان این آزمون‌ها را با دقت بیشتری انجام داده و دقت ایجاد تمایز بین شرکت‌کنندگان را بالا خواهد برد، اما دارای پیامدهایی نیز برای سازمان برگزارکننده آنها خواهد بود.

---

[1]. Test of Language by the Iranian Measurement Organization-TOLIMO

مکانیسم گمشدگی

دلایل متعددی می‌تواند به ایجاد داده گمشده منجر شود. روش‌های جای‌گذاری داده‌های گمشده به مکانیسم گمشدگی آنها وابسته است. بنابراین لازم است این دلایل در تحلیل مقادیر گمشده مد نظر قرار گیرند (فلیس و همکاران[1]، ۲۰۰۲). در عمل، امکان تشخیص نوع مکانیسم گمشدگی کار چندان آسانی نیست. گاهی محققان به‌ناچار با فرض پذیرش مکانیسم مورد نظر، به مطالعه و تحلیل نتایج می‌پردازند. چهار نوع مکانیسم گمشدگی وجود دارد که در ادامه به صورت مختصر بیان می‌شوند.

۱. گمشدگی کاملا تصادفی[2] (MCAR): در این حالت مقادیر گمشده به سایر متغیرها وابسته نیستند (فلیس و همکاران، ۲۰۰۲)؛ به عبارت دیگر احتمال گمشدگی نه به قسمت مشاهده شده و نه به قسمت مشاهده نشده، وابسته نباشد. در این حالت احتمال گمشدگی برای تمام رکوردها یکسان است. به‌عنوان مثال اگر در یک پژوهش میدانی زمینه‌یابی، آزمودنی برای پاسخ به سؤال «درآمد» به‌واسطه نتیجه پرتاب یک طاس تصمیم بگیرد و اگر عدد «٦» ظاهر شد، از پاسخ دادن امتناع کند، در این صورت گمشدگی از نوع کاملا تصادفی است و در چنین شرایطی حذف آزمودنی‌های دارای مقادیر گمشده در استنباط آماری خللی وارد نمی‌کند.

۲. گمشدگی تصادفی[3] (MAR): در این حالت مقادیر گمشده به وضعیت متغیرهای مشاهده شده وابسته است (مارووالا[4]، ۲۰۰۹). به‌عنوان مثال اگر گمشدگی درآمد با توجه به سن اشخاص تغییر کند (مثلا گمشدگی درآمد، بیشتر در سنین بالاتر باشد) اما با توجه به مقدار خود درآمد در یک سن خاص تغییر نکند (مثلا گمشدگی در یک بازه سنی خاص الگویی تصادفی داشته باشد) گمشدگی از نوع تصادفی است (تمپل[5]، ۲۰۰۸).

۳. گمشدگی غیر تصادفی[6] (MNAR): در این حالت، مقادیر گمشده به وضعیت سایر متغیرهای گمشده وابسته است. به‌عنوان مثال مطالعه‌ای را در نظر بگیرید که در آن هدف اندازه‌گیری کیفیت زندگی افراد است (متغیر پاسخ). اگر افرادی که از سطح کیفیت بالا یا پایین برخوردارند، از پر کردن کامل پرسشنامه

---

1. Fleiss et al
2. Missing completely at random
3. Missing at random
4. Marwala
5. Templ
6. Missing not at random

خودداری کنند، گمشدگی غیر تصادفی رخ داده است (زائری و همکاران، ۱۳۹۱). به‌عنوان مثالی دیگر؛ می‌توان گفت که مردم «تندخو و گستاخ» کمتر احتمال دارد به سؤال درآمد پاسخ دهند.

٤. گمشدگی که به ماهیت خود مقادیر گمشده بستگی دارد[1] (MDMVI): گمشدگی که در آن مقادیر به علت ذات پژوهش که در آن نمی‌توان آن مقادیر را به‌طور معمول اندازه‌گیری کرد، گمشده محسوب می‌شوند (فلیس و همکاران، ۲۰۰۲؛ ماروالا، ۲۰۰۹). به‌عنوان مثال، افراد با درآمدهای بالاتر کمتر احتمال دارد که درآمدشان را آشکار کنند.

نیرلی و همکاران[2] (۲۰۰۵) نشان داده‌اند که MCAR، MAR و NMAR معمولاً منجر به برآوردهای متفاوتی برای یک روش جانهی می‌شوند و روش‌های جانهی متفاوت نیز معمولا برآوردهای متفاوتی را برای مکانیسم گمشدگی مورد نظر ارائه می‌دهند. تنها مکانیسمی که همواره ضریب رگرسیونی را برای تمام روش‌های جانهی کاهش می‌دهد، مکانیسم MAR است. نتایج بررسی‌های قبلی نشان داده است که خطای استاندارد در ۲۵ درصد از روش‌های جانهی تحت MCAR، در ٤۲ درصد از روش‌های جانهی تحت MAR و در ۵۰ درصد از روش‌های جانهی تحت NMAR افزایش می‌یابد. همچنین می‌توان گفت که بهترین روش برای برآورد داده داده‌های دقیق‌تر و همچنین اندازه‌گیری اریبی در برآوردها، داشتن دانش پیشین کافی از برآوردها و ارتباط بین متغیرهای پاسخ و پیش‌بین قبل از شروع تحلیل اصلی است. درنتیجه، در این حالت باید تلاش منظم و برنامه‌ریزی شده‌ای اتخاذ کرد تا سطوح داده‌های گمشده کاهش یابد و همچنین متغیرهایی جمع‌آوری کرد که همبستگی بالایی با متغیرهای پاسخ داشته باشند و تا حد ممکن سعی کرد درباره چرایی داده‌های گمشده، اطلاعات قبلی و جانبی جمع‌آوری شود.

برخی از روش‌های برخورد با داده‌های گمشده

۱. روش حذف داده گمشده[3]: این روش در گذشته مرسوم‌ترین روش برخورد با داده گمشده بوده است. اما استفاده از آن موجب کاهش حجم نمونه شده و باعث می‌شود که برآورد پارامترها اریب شود (گلین و لارد[4]، ۱۹۸۳). همچنین در این روش

---

1. Missingness that depends on the missing value itself
2. Nirelli et al
3. Deletion of missing data
4. Glynn & Laird

حذف مقادیر گمشده می‌تواند به دور ریختن اطلاعات با ارزش منجر شود و داده‌های باقیمانده نمونه خوبی برای کل داده‌ها نباشند (وایمن، ۲۰۰۳).

۲. الگوریتم EM[1]: این الگوریتم را در اواخر دهه ۱۹۷۰ روبین، دمپستر و لارد معرفی کردند و توسعه دادند (دمپستر و همکاران[2]، ۱۹۷۷). با توجه به اینکه در مقادیر مشاهده شده اطلاعاتی در خصوص احتمال مقادیر گمشده وجود دارد، این الگوریتم از سایر متغیرها برای جایگذاری مقدار گمشده در یک متغیر استفاده می‌کند[3] و بررسی می‌کند که آیا این مقدار محتمل‌ترین مقدار است[4] و اگر نباشد مقدار دیگری جایگزین می‌شود و این روند تا رسیدن به محتمل‌ترین مقدار ادامه پیدا می‌کند. انتخاب نام EM نیز به علت یک مرحله امید ریاضی‌گیری و سپس ماکسیمم‌سازی در تکرار الگوریتم است (دمپستر و همکاران، ۱۹۷۷). این الگوریتم از داده‌های کامل برای محاسبه میانگین، واریانس و کواریانس استفاده می‌کند. پس از آن برای به دست آوردن خطوط رگرسیون ارتباط هر متغیر به سایر متغیرها، روش ماکسیمم درست‌نمایی[5] (ML) به‌کار می‌رود. در این مرحله به تعداد متغیرها، معادله خواهیم داشت. ML این اطمینان را به ما می‌دهد که این معادله‌ها دقیق‌ترین میانگین، واریانس و کواریانس را ارائه می‌دهند. با استفاده از این معادلات، مقادیر گمشده برآورد می‌شوند و مجموعه داده‌های ما در این مرحله کامل می‌شود. سپس با استفاده از این مجموعه داده کامل، دوباره میانگین، واریانس و کواریانس برآورد می‌شوند که ممکن است با مقادیر قبلی کمی متفاوت باشند؛ چراکه در این مرحله با استفاده از مجموعه داده‌های کامل برآورد شده‌اند. دوباره معادلات رگرسیون با استفاده از ML محاسبه می‌شوند و مجدداً مقادیر برآورد می‌شوند. این سه مرحله تا رسیدن به همگرایی تکرار می‌شوند. با جایگذاری مقادیر گمشده با استفاده از EM ارتباط بین متغیرها حفظ می‌شود (روبین و روتنیزکی[6]، ۱۹۹۲). یکی از پیچیدگی‌های روش الگوریتم EM آن است که نیازمند مدل‌بندی پارامترهای مزاحم[7] متغیرهای کمکی است. در برخی مواقع فقط با تعداد کمی از متغیرهای طبقه‌ای توزیع چندجمله‌ای اشباع شده[8] می‌تواند برازش داده شود. زمانی که

1. Expectation Maximization
2. Dempster et al
3. Expectation
4. Maximization
5. Maximum likelihood
6. Robins & Rotnitzky
7. nuisance parameters
8. Polynomial distribution saturated

متغیرهای بیشتری وجود دارند، اغلب انجام برخی ساده‌سازی‌ها برای توزیع توأم[۱] ضروری است (هورتون و همکاران[۲]، ۲۰۰۷).

۳. الگوریتم داده‌افزایی[۳]: این روش نیز مانند الگوریتم EM محاسبه‌ای تکراری است که به‌طور متناوب داده‌های گمشده را جای‌گذاری می‌کند و سپس پارامترهای ناشناخته را با روندی تصادفی پیش‌بینی می‌کند. در این روش نخست جایگذاری ابتدایی برای داده‌های گمشده بر اساس مقادیر فرضی پارامترها در نظر گرفته می‌شود سپس پارامترهای جدید با استفاده از توزیع پسین به دست آمده از داده‌های کامل برآورد می‌شود. فرایند شبیه‌سازی پارامترها و داده‌های گمشده یک زنجیر مارکف[٤] تولید می‌کند که سرانجام تثبیت شده یا در توزیع همگرا[٥] می‌شود (آشفته، ۱۳۹۲). این روش می‌تواند با تکرار مراحل زیر حاصل شود (تنر و وانگ[٦]، ۱۹۸۷؛ دُنی و همکاران[۷]، ۲۰۱۵).

گام Iام جانهی: با بردار میانگین و ماتریس کوواریانس برآورد شده، مقادیر گمشده برای هر مشاهده به‌طور مستقل شبیه‌سازی می‌شوند. به این معنی که اگر شما متغیرهای با مقادیر گمشده را برای مشاهده iام با $Y_{i(mis)}$ نشان دهید و متغیرهای با مقادیر مشاهده شده را با $Y_{i(obs)}$، بنابراین مرحله اول، مقادیر را برای $Y_{i(mis)}$ از توزیع شرطی $Y_{i(mis)}|Y_{i(obs)}$ تولید می‌کند.

گام pام پسین: بردار میانگین و ماتریس کوواریانس پسین جامعه از برآوردهای نمونه کامل شبیه‌سازی شده‌اند. سپس این برآوردهای جدید در گام I استفاده می شوند. بدون اطلاع قبلی درباره پارامترها، یک پیشین ناآگاهی بخش استفاده می‌شود.

این دو مرحله به اندازه کافی برای نتایج قابل اعتماد برای مجموعه داده‌های جانهی چندگانه مؤثر هستند (شافر[۸]، ۱۹۹۷). اغلب تعداد کمی از جانهی‌ها در جانهی چندگانه مناسب هستند (روبین، ۱۹۹٦).

---

1. Joint distribution
2. Horton et al
3. Data augmentation
4. Markov chains
5. Convergent
6. Tanner and Wong
7. Donneau
8. Schafer

٤. روش جانهی چندگانه[1]: روش جانهی چندگانه یا جانهی بیشتر از یک مقدار جانهی شده به داده‌های گمشده را روبین در سال ۱۹۷۸ معرفی کرد و پس از آن، توسط روبین در سال‌های ۱۹۸۷ و ۱۹۹٦ گسترش یافت. در این روش هر مقدار گمشده با مجموعه مقادیر به دست آمده از توزیع پیش‌بینی شده جایگذاری می‌شوند (روبین، ۱۹۸۷). این روش رویکردی مبتنی بر شبیه‌سازی برای تحلیل آماری داده‌های ناکامل[2] است. در جانهی چندگانه هر داده گمشده توسط m>۱ مقدار شبیه‌سازی شده، جایگذاری می‌شود. به عبارتی، m نسخه از داده‌های کامل را می‌توان با استفاده از روش‌های استاندارد داده‌های کامل تجزیه و تحلیل و نتایج آنها را به منظور استنباط آماری ترکیب کرد (به‌عنوان مثال برای ایجاد برآوردهای فاصله‌ای یا مقادیر p) و از این راه، مسئله نامشخص بودن داده‌های گمشده را حل کرد (شافر، ۱۹۹۷). ازجمله مزایای این روش، امکان محاسبه چندین بار برآوردهای خطای معیار با استفاده از تکنیکی یکسان است (لیتل و روبین، ۱۹۸۷).

برخی از مزایا و معایب هر یک از روش‌های ذکر شده به‌طور مختصر در جدول (۱) ارائه شده است.

جدول (۱) مزایا و معایب روش‌های جانهی

ردیف	روش جانهی	مزایا	معایب
۱	حذف فهرستی	۱. سادگی ۲. قابلیت مقایسه بین تحلیل‌های مختلف	۱. کاهش قدرت آماری (به علت کاهش حجم داده‌ها) ۲. استفاده نکردن از تمام داده‌ها ۳. اگر داده‌ها MCAR نباشند ممکن است برآوردها اریب‌دار شوند.
۲	حذف جفتی	۱. تمام موارد ممکن را برای تجزیه و تحلیل نگه می‌دارد. ۲. تمام اطلاعات ممکن را با هر تجزیه و تحلیل مورد استفاده قرار می‌دهد.	۱. از آنجا که نمونه‌ها در تحلیل‌های مختلف یکسان نیست، نتایج تحلیل‌ها را نمی‌توان مقایسه کرد.
۳	جانهی میانگین	۱. می‌توان تمام روش‌های تحلیل داده‌های کامل را به‌کار برد.	۱. کاهش تغییرپذیری ۲. به دلیل نادیده گرفتن ارتباط بین متغیرها باعث ایجاد اریبی در برآوردهای مربوط به تحلیل‌های کوواریانس و

[1]. Multiple imputation
[2]. incomplete data

ردیف	روش جانهی	مزایا	معایب
			همبستگی می‌شود
٤	الگوریتم EM	۱. از اطلاعات کامل (در هر دو موقعیت داده کامل و ناقص) را برای محاسبه لگاریتم درست‌نمایی استفاده می‌کند. ۲. شناسایی مجموعه مقادیری از پارامترها که بالاترین لگاریتم درست نمایی را تولید می‌کند. ۳. با برقراری فرض‌های MCAR و MAR در داده‌ها، برآوردهای نااریبی از پارامترها انجام می‌دهد (لیتل و روبین، ۱۹۸۷). ٤. این الگوریتم معمولاً به‌عنوان نقطه شروعی برای روش جانهی چندگانه استفاده می‌شود.	۱. خطاهای استاندارد با تمایل به سمت کم برآورد شدن با استفاده از ماتریس اطلاعات مشاهده شده می‌توانند تعدیل شوند. ۲. خطای استاندارد ناریب برآورد پارامترها را فراهم می‌کند (سالکیند و راسموسن [1]، ۲۰۰۷). ۳- معمولا به‌عنوان روشی برای تولید مقادیر اولیه برای داده‌های گمشده، در روش‌های دیگر مانند جانهی چندگانه یا داده افزایی از این الگوریتم استفاده می شود و به تنهایی استفاده چندانی در تحلیل‌های پیشرفته ندارد و به برآوردهای اریب منجر می‌شود.
٥	الگوریتم داده افزایی	۱. در نظر گرفتن مقادیر گمشده به‌عنوان پارامتر و برآورد آنها با استفاده از روش‌های مونت‌کارلوی زنجیر مارکوفی (تنر و وانگ، ۱۹۸۷) ۲. در نظر گرفتن عدم قطعیت موجود در داده‌های گمشده ۳. هنگامی که داده‌های گمشده دارای الگوی گمشدگی پیچیده باشند استفاده از این روش نسبت به سایر روش‌ها کاراتر و انجام‌پذیرتر است.	۱. نیاز به محاسبات بیشتر نسبت به روش هایی مانند جانهی و حذف دارد.
٦	جانهی چندگانه	۱. تغییرپذیری دقیق‌تر با جانهی‌های چندگانه برای هر مقدار گمشده (تغییرپذیری ناشی از نمونه‌گیری و جانهی را در نظر می‌گیرد). ۲. خطای استاندارد برآورد پارامترها با استفاده از این روش نااریب باقی	۱. برنامه‌نویسی سخت و طاقت‌فرسا ۲. امکان بروز خطا هنگام مشخص کردن مدل ۳. تولید برآوردهای اریب‌دار (لیتل و روبین، ۱۹۸۷)

[1]. salkind & Rasmussen

ردیف	روش جانهی	مزایا	معایب
		می‌ماند (روبین، ۱۹۸۷) ۳. استفاده از رگرسیون تصادفی در این روش می‌تواند از کم برآوردی برآوردها پیشگیری کند.	

در ادامه نرم‌افزارهای مورد استفاده در سه روش الگوریتم EM، الگوریتم داده‌افزایی و جانهی چندگانه در برخورد با داده‌های گمشده شرح داده می‌شوند.

نرم‌افزارهای کاربردی برای استفاده از الگوریتم EM، الگوریتم داده‌افزایی و جانهی چندگانه در برخورد با داده‌های گمشده

الف) الگوریتم EM: این الگوریتم برای برآورد ML داده‌های گمشده با استفاده از ماتریس کواریانس بدون ساختار در چندین برنامه در دسترس است. اولین پیاده پیاده‌سازی تجاری توسط BMDP (نرم‌افزار آماری BMDP، ۱۹۹۲) منتشر شد و در حال حاضر در منوی داده گمشده SPSS گنجانده شده است. این روش همچنین در EMCOV (گراهام و هافر[1]، ۱۹۹۱)، SAS (یان[2]، ۲۰۰۰)، S-PLUS (شیمرت[3] و همکاران، ۲۰۰۱)، LISREL (جورسکوگ و سوربن[4]، ۲۰۰۱) و Mplus (موسن و موسن[5]، 1998) قابل دسترس است. ML همچنین برای مدل‌های نرمال با ماتریس کواریانس ساختار یافته در دسترس است. مدل‌های خطی چندسطحی می‌توانند با HLM (بریک[6] و همکاران، ۱۹۹۶)، MLwin (پروژه مدل‌های چندسطحی[7]، ۱۹۹۶)، روش PROC MIXED در SAS (لیتل و همکاران، ۱۹۹۶)، Stata (Stata، ۲۰۰۱) و تابع lme در S-PLUS (اینسایتفول[8]، ۲۰۰۱) برازش داده شوند. هر کدام از اینها ممکن است برای داده‌های اندازه‌گیری مکرر نیز استفاده شوند. در برخی موارد، این روش‌ها صرفاً در پیشینه پژوهشی متون مقادیر گمشده ذکر نمی شوند بلکه با شرایطی مجموعه داده‌های «نامتعادل» را توصیف می‌کنند، که در آن داده‌ها در یک زمان واحد اندازه‌گیری نشده است و نوعی حالت عدم تعادل زمانی در

1. Graham & Hofer
2. Yuan
3. Schimert
4. Jo¨reskog & So¨rbom
5. Muthe´n & Muthe´n
6. Bryk
7. Multilevel Models Project
8. Insightful

اندازه‌گیری داده‌ها وجود دارد. در اینجا باید تأکید شود که اگر عدم تعادل زمانی رخ دهد اما نه به دلیل ماهیت طرح پژوهشی بلکه به دلیل بی‌پاسخی کنترل نشده (مثل افت نمونه‌گیری)، همه این برنامه‌ها فرض را بر MAR می‌گذارند. برآوردهای ML برای مدل‌های معادلات ساختاری با داده‌های ناکامل در MX (نیل[۱] و همکاران، ۱۹۹۹)، AMOS (آربوکل و ووسک[۲]، ۱۹۹۹)، LISREL و Mplus که هر کدام فرض MAR را در نظر می‌گیرند، قابل دسترس است. این برنامه‌ها خطای استاندارد را بر اساس اطلاعات مورد انتظار یا مشاهده شده فراهم می‌کنند. اگر یک انتخاب پیشنهاد شود، کاربر باید مقادیر مشاهده شده را به جای مقادیر مورد انتظار انتخاب کند، زیرا برنامه آخری فقط تحت MCAR مناسب است.

پیچیدگی دیگر برای ماکسیمم درست‌نمایی، مرتبط با محاسبه خطای استاندارد برآوردهاست. اجرای ماکسیمم درست‌نمایی که اشاره به این پیچیدگی دارد در نسخه ۷ LogXact، بخش تحلیل مقادیر گمشده S-plus و SPSS قابل دسترس است (ون هیپل[۳]، ۲۰۰٤).

نرم‌افزار Amelia II (هوناکِر[٤] و همکاران، ۲۰۰٦) گام‌های جانهی را انجام می دهد و از الگوریتم EM بر اساس بازنمونه‌گیری استفاده می‌کند (کینگ و همکاران[٥]، ۲۰۰۱) که دقت و سرعت عمل بالایی دارد و شامل ویژگی‌هایی برای جانهی نظرسنجی مقطعی[٦]، داده‌های سری زمانی[۷] و داده‌های سری زمانی/مقطعی[۸] است. این بسته نرم‌افزاری زمانی که اطلاعات مورد نیاز در دسترس است، امکان قرار دادن مقادیر پیشین را برای خانه‌های گمشده فردی در ماتریس داده‌ها فراهم می‌کند. هر کدام از تحلیل‌ها به صورت جداگانه و ترکیبی می‌توانند تحت R با استفاده از نرم‌افزار Zelig (ایما[۹] و همکاران، ۲۰۰٦)، یا یک بسته آماری جداگانه (به‌عنوان مثال SAS یا Stata) انجام شوند. مقاله هوناکِر و کینگ (۲۰۰٦) توضیح‌های بیشتری را در خصوص این بسته ارائه می‌دهد.

---

1. Neale
2. Arbuckle & Wothke
3. von Hippel
4. Honaker
5. King et al
6. Sectional surveys
7. Time series data
8. Time series data / cross
9. Imai

علاوه بر این، یک بسته نرم‌افزاری تکمیلی در دسترس است که نصب Amelia II را بدون نیاز به هر دانشی و یا حتی اجرای مستقیم، در سیستم R به کاربر اجازه می‌دهد، اگر این مسیر ترجیح داده شود، Amelia می‌تواند مجموعه داده را برای تحلیل و ترکیب در بسته دیگری، خروجی دهد (هورتون و همکاران، ۲۰۰۷).

ب) **الگوریتم داده‌افزایی**: نسخه ۷ S-Plus بخش مربوط به داده‌های گمشده را که S-Plus برای پشتیبانی مدل بر اساس مدل‌های داده‌های گمشده با استفاده از روش‌های شافر (۱۹۹۷) توسعه داده است را با استفاده از الگوریتم EM (دِمپستر و همکاران[1]، ۱۹۷۷) و الگوریتم داده‌افزایی (DA) (تنِر و همکاران، ۱۹۸۷) نشان می دهد. الگوریتم DA می‌تواند برای تولید جانهی‌های چندگانه استفاده شود. بخش مربوط به داده‌های گمشده یک پشتیبانی را برای داده‌های نرمال چندمتغیره[2] (imp Gauss)، داده‌های طبقه‌ای[3] (impLoglin) و مدل‌های گاوسی شرطی[4] (impcgm) برای جانهی‌هایی شامل هر دو متغیرهای گسسته و پیوسته فراهم می‌کند.

ج) جانهی چندگانه: برنامه جانبی S+MissingData در نرم‌افزار S-PLUS تعداد بسیار زیادی از روش‌های تحلیل داده‌های گمشده شرح داده شده در شافر (۱۹۹۷) را ارائه کرده است. این برنامه مجموعه‌ای از توابع را برای برازش چندمتغیره گاوسی[5]، لگاریتم خطی[6] و مدل‌های عمومی موضعی[7] با استفاده از الگوریتم EM و الگوریتم‌های داده افزایی (DA) در اختیار دارد. الگوریتم‌های DA جانهی چندگانه را نیز اجرا می‌کنند. در این نرم‌افزار مجموعه مربوطه بر اساس کد شافر ساخته شده است، اما در برخی موارد الگوریتم‌های متفاوتی را نیز استفاده کرده است. به‌عنوان مثال الگوریتم EM برای برازش مدل گاوسی، یک تجزیه چولسکی[8] از کواریانس را به جای روفتن[9] در تابع (imp.norm) شافر، استفاده می‌کند (شافر، ۱۹۹۷).

برخی از بسته‌های نرم‌افزاری درS-PLUS عبارتند از NORM (جانهی چندگانه داده‌های پیوسته چندمتغیره با یک مدل نرمال)،CAT (جانهی چندگانه داده‌های طبقه‌ای با مدل‌های لگاریتم خطی)، MIX (جانهی چندگانه داده‌های پیوسته و طبقه

---

1. Dempster et al
2. Multivariate normal
3. Categorical data
4. Conditional Gaussian models
5. multivariate Gaussian
6. log-linear
7. general location models
8. Cholesky decomposition
9. sweeps

ای آمیخته با مدل عمومی موضعی[1]) و PAN (جانهی چندگانه داده‌های پنلی یا داده-های خوشه‌ای با مدل اثرات آمیخته خطی چندمتغیره[2]) که به‌عنوان توابع در PLUS -S قابل دسترس هستند. برای توضیح بیشتر در خصوص این نرم‌افزارها به شافر ۱۹۹۷b و فصول ۵، ۷، ۸ و ۹ شافر ۱۹۹۷a مراجعه شود.

تجزیه و تحلیل روش جانهی چندگانه در SAS/STAT در سه مرحله انجام می شود. ابتدا، جانهی توسط PROC MI انجام می‌شود، سپس، روش داده‌های کامل با استفاده از هر روش SAS برای تحلیل داده‌های کامل مورد استفاده قرار می‌گیرد (به‌عنوان مثال، PHREG, GENMOD, PROC GLM یا LOGISTIC)؛ توسط دستور 'BY' تحلیل برای هر مجموعه داده کامل تکرار می‌شود. در نهایت، با استفاده از PROC MIANALYZE ترکیب می‌شوند. به هیچ نصب اضافه‌ای برای PROC MI/PROC MIANALYZE نیاز نیست، زیرا بخشی از نرم‌افزار SAS/STAT است (هورتون و همکاران، ۲۰۰۷).

در SPSS قسمتی از منوی مقادیر گمشده، جانهی چندگانه را توسط معادلات زنجیری[3] پشتیبانی می‌کند. جانهی و تحلیل اصلی داده‌ها می‌تواند در قالبی اتوماتیک و کاربرپسند انجام شود و این روش به‌خوبی با دیگر کاربردهای تحلیلی این نرم‌افزار برای تحلیل داده‌های کامل، یکپارچه و ادغام شده است (شافر، ۱۹۹۷).

روش

در پژوهش حاضر نتایج نود و یکمین آزمون تولیمو در بخش ساختار با ۵۰ سؤال چهارگزینه‌ای بررسی شد. تعداد کل شرکت‌کنندگان ۲۰۲۸ نفر بودند که از این تعداد، ۱۲۸۹ شرکت‌کننده به تمامی سؤال‌ها پاسخ داده بودند. با استفاده از نرم‌افزار R برای این افراد در هر ۵۰ متغیر بین ۷ تا ۱۰ درصد مقادیر گمشده تولید شد. سپس مقادیر گمشده در داده‌های ناکامل با سه روش الگوریتم داده افزایی، الگوریتم EM و جانهی چندگانه برآورد شدند و نمره کل شرکت‌کنندگان بر اساس داده‌های جدید، محاسبه شد. با در دست داشتن داده‌های کامل اولیه، معیار MSE برای هر سه روش محاسبه و با استفاده از آن به مقایسه سه روش فوق پرداخته شد. در ادامه نظریه به‌کار گرفته شده و مقدار MSE برای سه روش مذکور ارائه شده است.

---

1. Multiple imputation of mixed continuous and categorical data
2. Multivariate linear mixed effects
3. Chained equations

یافته‌ها

الگوریتم داده‌افزایی در نرم‌افزار R انجام گرفت. برای این منظور فرض کنید داده داده‌ها با اندازه n با انتخاب‌های p> ۲ باشد. در اینجا منظور از انتخاب‌ها تعداد رده‌ها در مدل چندجمله‌ای است. به‌عنوان مثال در آزمون تولیمو داوطلبان از بین چهار گزینه می‌توانند یکی را انتخاب کنند. این مدل، متفاوت از مدل پروبیت ترتیبی[1] است. در مدل چندجمله‌ای پروبیت، یک توزیع نرمال چند متغیره برای متغیرهای پنهان به‌صورت $w_i = (w_{i,1},...,w_{i,p-1})$ فرض می‌شود که

$$w_i = X_i\beta + e_i, \quad e_i \sim N(0, \sum), \quad i = 1,...,n$$

که X_i یک ماتریس $(p-1)\times k$ است، k تعداد متغیرهای کمکی (تبیینی) است و در اینجا فقط ازجمله عرض از مبدأ استفاده می‌کنیم. e_i یک بردار $(p-1)\times k$ و $\sum$ یک ماتریس همیشه مثبت $(p-1)\times(p-1)$ است. برای اینکه مدل شناسایی شناسایی‌پذیر[2] باشد، نخستین عنصر قطری $\sum$ به‌صورت $\sigma_{11} = 1$ مقید می‌شود. متغیر پاسخ y_i انتخاب فرد iام در گزینه‌ها است و برحسب این متغیر پنهان مدل می می‌شود،

$$y_i(w_i) = \begin{cases} 0 & if \quad \max(w_i) < 0 \\ j & if \quad \max(w_i) = w_{ij} > 0 \end{cases}$$

به ازای $i = 1,...,n \quad j = 1,..., p-1$، که y_i مساوی با صفر متناظر با رسته مرجع است.

توزیع پیشین برای مدل پروبیت چندجمله‌ای به‌صورت

$$\beta \sim N(0, A^{-1}),\ p(\sum) \propto \left|\sum\right|^{-(\nu+p)/2} \left[trace(S\sum^{-1})\right]^{\frac{\nu(p-1)}{2}}$$

[1]. Ordered probit model
[2]. Identify possible

که A ماتریس دقت پیشین β، ν درجه آزادی پیشین $\sum$ است و ماتریس $(p-1)\times(p-1)$ همیشه مثبت S، مقیاس پیشین برای $\sum$ است. فرض می‌کنیم که نخستین عنصر قطری S برابر با یک است. توجه کنید که در اینجا مدل ما به صورت

$$W_i = \beta_0 + e_i,\ e_i \sim N(0, \sum),\ i = 1,...,n$$

که در آن $\beta_0 \sim N(0, k^{-1})$ و k مقدار دقت پیشین β مثلا ۰/۰۰۱ در نظر گرفته شد و بر این اساس مجموعه داده‌های کامل حاصل شد. مقدار MSE در این روش برابر با ۸۰/۸۷ به دست آمد.

الگوریتم EM نیز با استفاده از نرم‌افزار R انجام شد. برای این منظور هر متغیر چندجمله‌ای در Amelia تعریف شد. برای یک متغیر چندجمله‌ای با p رسته، Amelia، p-۱ متغیر دوجمله‌ای تعریف می‌کند. این p-۱ متغیر جدید به‌عنوان متغیرهای دیگری در روش جانهی چندمتغیره در نظر گرفته می‌شوند و به صورت پیوسته جانهی می‌شوند. این جانهی‌های پیوسته به‌طور مناسب درون احتمال‌های هر p رسته ممکن مقیاس‌بندی می‌شوند، سپس یکی از این رسته‌ها استخراج خواهد شد که بر مبنای متغیر چندجمله‌ای p رسته‌ای ساخته شده است. یعنی رسته مورد نظر از یک توزیع چندجمله‌ای استخراج می‌شود، بنابراین تمام استخراج‌ها چندجمله‌ای هستند. چون Amelia یک متغیر چندجمله‌ای p رسته‌ای را به‌عنوان p-۱ متغیر در نظر می گیرد. باید توجه داشت که اگر متغیرهای چندجمله‌ای زیادی استفاده شود، تعداد پارامترها به سرعت زیاد خواهند شد. در اینجا مقدار MSE در این روش برابر با ۳۱/۴۷ به دست آمد.

روش جانهی چندگانه با استفاده از روش جانهی رگرسیون لجستیک چندجمله‌ای در نرم‌افزار Stata انجام شد. فرض کنید $x = (x_1,...,x_n)'$، شامل k رسته بوده (بدون از دست دادن کلیت مسئله، فرض کنید k=۱ رسته مرجع باشد) و از مدل لجستیک چندجمله‌ای زیر پیروی کند:

$$p(x_i = k|z_i) = \begin{cases} \frac{1}{1+\sum_{l=2}^{k} \exp(z_i' \beta_l)}, & if\ k = 1 \\ \frac{\exp(z_i' \beta_l)}{1+\sum_{l=2}^{k} \exp(z_i' \beta_l)}, & if\ k > 1 \end{cases} \quad (1)$$

که z_i بردار مقادیر پیشگو است که در پژوهش حاضر تنها β_0 یعنی عرض از مبدأ در مدل لحاظ شده است. X شامل مقادیر گمشده است که بایستی برآورد شوند. افراز $x = (x_0', x_m')$ را در نظر بگیرید که بردارهای $n_0 \times 1$ و $n_1 \times 1$ بردارهای شامل مشاهدات کامل و ناقص هستند. بر این مبنا، جانهی چندگانه به صورت زیر عمل می کند:

۱) مدل (۱) را به داده‌های مشاهده شده (x_0, z_0) برای به دست آوردن برآوردهای ماکسیمم درست‌نمایی برازش می‌دهد که z_0 برداری از یک‌ها است (زیرا در اینجا متغیر کمکی نداریم)، z_0 و واریانس نمونه‌گیری مجانبی آنها $\hat{U}$ به دست می‌آید.

۲) پارامترهای جدید β_* از تقریب نرمال نمونه بزرگ $N(\hat{\beta}_0, \hat{U})$ برای به دست آوردن توزیع پسین آن با فرض پیشین ناآگاهی بخش $p(\beta_0) \propto 0$ شبیه‌سازی می‌شود.

۳) یک مجموعه مقادیر جانهی شده x_m' با شبیه‌سازی از توزیع لجستیک چندجمله‌ای به دست می‌آید: یکی از k رسته به‌طور تصادفی به رسته گمشده i_m با استفاده از احتمال‌های تجمعی محاسبه شده از رابطه (۱) تخصیص می‌یابد با $\beta_l = \beta_{*l}$ و $z_i = z_{im}$.

٤) مراحل ۲ و ۳ را تا به دست آوردن M مجموعه مقادیر کامل $x_m^1, x_m^2, ..., x_m^M$ تکرار می‌کنیم.

مراحل ۲ و ۳ متناظر با همان استخراج‌های تصادفی از توزیع پیشگوی پسین داده داده‌های $p(x_m|z_0, z_0)$ هستند. در اینجا M=۳ در نظر گرفته شد و میانگین MSE

سه مجموعه داده کامل، به‌عنوان MSE روش جانهی چندگانه در نظر گرفته شد که این مقدار برابر با ٦٨/٨٤ به دست آمد.

میزان MSE به دست آمده از سه روش فوق، به‌عنوان معیار انتخاب روش مناسب برای جانهی در پژوهش حاضر استفاده شد. همان‌طور که ملاحظه می‌شود این مقدار برای الگوریتم EM نسبت به دو روش دیگر کمتر بوده و لذا روش فوق برای جانهی مقادیر گمشده در داده‌های پژوهش حاضر مناسب‌تر است. همان‌طور که پیش از این بیان شد با استفاده از روش فوق، از اطلاعات کامل (در هر دو موقعیت داده کامل و ناقص) برای محاسبه لگاریتم درست‌نمایی استفاده می‌شود و برآوردهای نااریبی از پارامترها را ارائه می‌دهد. همچنین روش الگوریتم EM نه‌تنها در حالت گمشدگی تصادفی که در حالت گمشدگی با احتمالات نابرابر نیز عملکرد خوبی از خود نشان می‌دهد (افشاری صفوی و همکاران، ١٣٩٤).

بحث و نتیجه‌گیری

از آنجا که جمع‌آوری داده‌های کامل در پژوهش‌های عملی چندان مقدور نیست، مقادیر گمشده اغلب در تمام پژوهش‌های علوم رفتاری، پزشکی و تحقیقات زمینه‌یابی وجود دارند و مشکلی اساسی در تجزیه و تحلیل این داده‌ها به‌شمار می‌روند. مقادیر گمشده، حجم اطلاعات را کاهش می‌دهد و موجب عدم تطابق نمونه و جامعه می‌شود. میزان این گمشدگی می‌تواند در نتایج به صورت متفاوت اثرگذار بوده و منجر به نتیجه‌گیری اشتباه شود و هرچه مقادیر گمشده افزایش یابد به تبع آن میزان اریبی برآوردها نیز افزایش می‌یابد. از طرفی استفاده از روش‌های برآورد پارامترها نیازمند داده‌های کامل است، لذا در هنگام برخورد با داده‌های ناکامل، استفاده از روش‌های مناسب جانهی ضروری می‌شود. در این خصوص، موضوع اساسی برای یک تحلیلگر مشخصات مناسب مدل جانهی است، چراکه نامشخص بودن این پارامترها می‌تواند دلیلی بالقوه برای اریبی باشد. اغلب در این خصوص یک مدل نرمال چندمتغیره، استفاده می‌شود زیرا از نظر محاسباتی انعطاف‌پذیر است (فقط بردار میانگین و ماتریس واریانس–کوواریانس نیاز به برآورد شدن دارند). این مدل حتی زمانی که برخی متغیرها دارای توزیع گاوسی نباشند می‌تواند استفاده شود؛ اگرچه این مسئله تحلیل‌ها را پیچیده کرده و اگر مقادیر جانهی شده گرد شده باشند، می‌تواند به اریبی منجر شود (هورتون و همکاران، ۲۰۰۳؛ آلیسون[1]، ۲۰۰۵؛ برنارد و همکاران[2]،

[1]. Allison
[2]. Bernaards et al

۲۰۰۷). این مسائل به‌ویژه زمانی که طبقه‌بندی چندگانه و متغیرهای پیوسته مقادیر گمشده دارند بیشتر قابل توجه می‌شود، زیرا توزیع توأم ممکن است پیچیده شود. در نهایت توجه داشته باشید که تجزیه و تحلیل‌ها نباید یک مدل غنی‌تر از مدلی را که برای جانهی به کار رفته است، مورد استفاده قرار دهند (لیتل و روبین، ۲۰۰۲).

هایتووسکی[1] (۱۹٦۸) عملکرد حذف فهرستی (معروف به روش کلاسیک) و حذف جفتی در زمینه رگرسیون خطی را بر اساس هشت نمونه کامل با حجم n=۱۰۰۰ شبیه‌سازی کرد که قسمتی از داده‌های این نمونه‌ها دارای مقادیر گمشده بودند. این هشت نمونه با توجه به تعداد کل متغیرها، توزیع متغیرهای پیش‌بین، ماتریس واریانس-کوواریانس و متغیرهای وابسته مرتبط با تغییرات عبارات خطا، از یکدیگر متفاوت بودند. هایتووسکی با مقایسه بین برآورد پارامترهای رگرسیونی به دست آمده از دو روش بررسی بر اساس نمونه‌های کاهش یافته با برآورد پارامترهای نمونه‌های کامل، نشان داد که حذف فهرستی تحت تمام شرایط به‌جز زمانی که نسبت داده‌های گمشده خیلی زیاد باشد (بیش از ۰,۹) یا زمانی که داده‌ها در الگویی به‌شدت غیر تصادفی گمشده باشند، بهترین عملکرد را خواهد داشت.

همچنین در خصوص تعیین کمی اثر اندازه نمونه و نسبت داده گمشده بر عملکرد روش جانهی مقادیر گمشده مطالعات اندکی صورت پذیرفته است که در این زمینه هایتووسکی (۱۹٦۸) اشاره کرده است که جانهی میانگین در رگرسیون خطی می‌تواند در برآورد پارامترها به‌طور جدی اریبی ایجاد کند. دلیل اصلی این اریبی آن است حتی اگر میانگین کلی با جانهی میانگین تغییر نکند، خطای استاندارد میانگین می‌تواند به‌طور قابل توجهی بسته به نسبت داده‌های گمشده کوچک‌تر شود. برای متغیر Y با n مشاهده، که k داده گمشده با میانگین n-k مشاهده غیر گمشده جایگزین شده‌اند، مجذور خطای استاندارد میانگین می‌تواند به صورت زیر نشان داده شود:

$$SE_M^2 = \frac{\sum_{i=1}^{n-k}(X_i - M)^2 + \sum_{i=n-k+1}^{n}(X_i - M)^2}{n(n-1)}.$$

رابطه بین نسبت داده‌های گمشده و کارایی روش برآورد داده‌های گمشده که توسط روبین (۱۹۸۷) ارائه شد، بسیار مهم است. زیرا این مسئله نشان می‌دهد که برای n بزرگ، افزایش نسبت داده‌های گمشده می‌تواند با افزایش در تعداد کل جانهی‌های چندگانه جبران شود. بنابراین اصل این مسئله بستگی به دیدگاه و نظر

[1]. Haitovsky

خود پژوهشگر دارد که تعیین کند که در قبال پیچیده‌تر شدن روش محاسباتی برآورد داده‌های گمشده، به چه میزان کارایی نیاز دارد. درواقع، روش جانهی چندگانه به خودی خود هیچ محدودیتی را تحمیل نمی‌کند.

گراهام و همکاران (۱۹۹٦) داده‌های شبیه‌سازی شده را برای ارزیابی روش‌های بررسی داده‌های گمشده شامل حذف جفتی، جانهی میانگین، جانهی تصادفی مفرد، جانهی چندگانه و جانهی چندگانه تغییرات ماکسیمم درست‌نمایی را در زمینه تحقیقات رفتاری، مورد استفاده قرار دادند. یافته‌های آنها نشان داد که با فرضیه MCAR، روش‌های ماکسیمم درست‌نمایی و جانهی چندگانه بهتر از حذف جفتی که به‌نوبه خود برتر از جانهی میانگین است، عمل می‌کند. با این حال، به‌استثنای روش ماکسیمم درست‌نمایی، برآورد پارامترها در تمام روش‌ها با برقراری فرض MCAR، اریب هستند. آنها در مطالعه خود با یک نمونه به حجم ۱۹٤۵ و درصد داده‌های گمشده ۵,۷٪ و ۱۱,٦٪، نشان دادند که افزایش نسبت داده‌های گمشده اریبی بزرگ‌تری در برآورد ایجاد می‌کند. در یک مطالعه مشابه، وایمن (۲۰۰۳)، ۱۹۳۷۳ مورد از یک ارزیابی آزمون خواندن ملی را مورد استفاده قرار داد که حدوداً ۱۵٪ داده ها دارای داده گمشده و چهار متغیر برای مقایسه عملکرد حذف فهرستی، جانهی میانگین و جانهی چندگانه وجود داشت. بر اساس میانگین‌های نمونه و خطای استاندارد آنها برای یک آزمون دارای نمره استاندارد شده، او به این نتیجه رسید که جانهی چندگانه بهترین عملکرد را دارد و پس از آن حذف فهرستی و جانهی میانگین در رتبه‌های بعدی قرار دارند. با این حال آنها در مطالعه خود اثر تغییر در اندازه نمونه، نسبت داده‌های گمشده و روش تجزیه و تحلیل را در نظر نگرفتند.

پگ و اندرس[1] (۲۰۰٤) نشان دادند که روش‌های پیشرفته مانند جانهی ماکسیمم درست‌نمایی و جانهی چندگانه زمانی که داده‌ها مبتنی بر فرض MAR باشند نسبت به حذف فهرستی عملکرد بهتری دارند.

اگرچه جانهی چندگانه در نمونه‌هایی با حجم بالا خیلی خوب عمل می‌کند، اما در نمونه‌های کوچک، برآوردهای اریب‌دار ایجاد می‌کند. برای مثال کیم[2] (۲۰۰٤) مقدار دقیق این نوع برآورد اریب‌دار را با استفاده از شبیه‌سازی مونت‌کارلو با ۵۰۰۰۰ نمونه و ۵ جانهی برای ۲ ×۳ ×۲ طرح عاملی محاسبه کرده است. او نشان داد زمانی که اندازه نمونه از ۲۰۰ به ۲۰ کاهش می‌یابد، واریانس برآورد پارامترهای جانهی چندگانه

---

1. Peugh & Enders
2. Kim

توسط یک عامل از ۱۰ عامل یا بیشتر، با دامنه نسبت گمشدگی بین ۰/۲ و ۰/٦، می می‌تواند افزایش یابد. نتایج این مطالعه بر اساس جانهی چندگانه با خواص آماری مطلوب‌تر از روش جانهی روبین (۱۹۸۷) یک روش جدید جانهی داده گمشده را برای مورد خاصی از اندازه نمونه بسیار کوچک (n≤۲۰) پیشنهاد کرد.

پنگ[1] و همکاران (۲۰۰٦) در مطالعه‌ای در حوزه تعلیم و تربیت، عملکرد دو ماکسیمم درست‌نمایی (اطلاعات کامل و ماکسیمم‌سازی امید ریاضی) و جانهی چندگانه را با حذف فهرستی با استفاده از دو نمونه واقعی با حجم ۱۳۰۲ و ۵۱۷ در زمینه تحلیل مسیر و رگرسیون لجستیک مقایسه کردند. آنها گزارش کردند که اندازه‌ها و یا نشانه‌های برآورد پارامترها، مقادیر p در آزمون فرضیه و توان آزمون می‌توانند به‌طور قابل توجهی بسته به روش داده گمشده متفاوت باشند و با فرض MAR، روش‌های جانهی ماکسیمم درست‌نمایی و جانهی چندگانه نسبت به حذف فهرستی برتر هستند. متأسفانه، پنگ و همکاران (۲۰۰٦) نمونه‌های با اندازه‌های متفاوت را با روش‌های متفاوت تجزیه و تحلیل بر اساس روش‌های بررسی داده گمشده، استفاده کردند. به همین دلیل، روابط متقابل بین این سه عامل برای این مطالعه نمی‌توانند ارزیابی شوند.

یانگ و همکاران[2] (۲۰۱۱) خلاصه‌ای از توصیه‌های مطالعات مختلف را برای فراهم کردن دستورالعمل‌های زیر ارائه کردند: ۱- زمانی که کمتر از ۱ درصد داده‌ها گمشده هستند، اثر روش‌های بررسی داده‌های گمشده بی‌اهمیت است؛ ۲- برای ۱ تا ۵ درصد داده گمشده، روش‌های ساده مانند حذف فهرستی و جانهی رگرسیون خوب عمل می‌کند؛ ۳- برای ۵ تا ۱۵ درصد داده گمشده، روش‌های پیچیده مانند جانهی چندگانه باید انتخاب شوند؛ و ٤-زمانی که داده گمشده بیش از ۱۵ درصد است، نتایج جانهی صرف نظر از روش جانهی استفاده شده تا حدود زیادی بی‌معنا هستند، زیرا خیلی کم می‌توانند درباره مکانیسم گمشدگی داده‌ها اطلاعات درستی ارائه کنند و برآوردگر حاصل اریب‌دار است. این نویسندگان معتقدند که تعداد بسیار محدودی از مطالعات در مورد افزایش توان به‌عنوان نتیجه مستقیم پیامد به‌کارگیری روش جانهی بحث کرده‌اند و در این خصوص انجام تحقیقات بیشتر را توصیه کرده‌اند تا نتایجی به دست آید که بتواند برای انتخاب بهترین روش جانهی مورد استفاده قرار گیرند. در نهایت، این نویسندگان پیشنهاد کرده‌اند که اگرچه جانهی چندگانه ممکن است در

---

1. Peng
2. Young et al

تمام شرایط بهترین نباشد، اما این روش به‌طور کلی در بیشتر مواقع بهترین روش یا دومین روش خوب است. حتی زمانی که دومین روش انتخابی است، تفاوت نسبی در عملکرد آن نسبت به بهترین روش در حداقل میزان ممکن است. به همین دلیل، آنها به‌طور کلی روش جانهی چندگانه را انتخابی مطمئن برای روش جانهی می‌دانند. این روش حتی در مواردی که نسبت داده‌های گمشده بزرگ است، عملکرد قابل قبولی ارائه می‌دهد.

با عنایت به مطالب مذکور و با توجه به اینکه میزان مقادیر گمشده در پژوهش حاضر برای هر متغیر بین ۷ تا ۱۰ درصد بود، از روش‌های پیچیده‌تر مانند الگوریتم EM، الگوریتم داده‌افزایی و جانهی چندگانه برای برآورد مقادیر گمشده، استفاده شد. این روش‌ها با لحاظ کردن ساختار داده‌ها و در نظر گرفتن توزیع‌های مناسب اعمال و در نهایت با توجه به میزان MSE آنها برای نمره داوطلبان در سؤال‌های مربوط به بخش ساختار مقایسه شدند. با مقایسه این روش‌ها بر اساس مقدار MSE، روشن شد که الگوریتم EM نسبت به سایر روش‌ها عملکرد بهتری دارد. رشیدی‌نژاد و نواب‌پور (۱۳۸۹) نیز در پژوهش خود با عنوان «مقایسه جانهی الگوریتم EM با دو روش جانهی میانگین و نمونه‌های جدید در آمارگیری‌های پانلی» نشان دادند که جانهی مقادیر گمشده با الگوریتم EM نسبت به جانهی با میانگین مشاهده‌های مشابه و روش جانهی با نمونه جدید کاراتر است.

همچنین در پژوهش پورحسینقلی و همکاران (۱۳۸٤) با عنوان «تحلیل درست‌نمایی ماکسیمم مدل رگرسیون لجستیک در حالتی که داده‌های متغیرهای پیشگو کامل نیستند ولی متغیرهای کمکی وجود دارند» نتایج حاصل از برآورد انحراف معیارهای دو روش تحلیل مورد کامل و جانهی با الگوریتم EM نشان داد که انحراف معیارهای برآورد شده برای تمام پارامترها در روش جانهی با الگوریتم EM کمتر از انحراف معیارهایی هستند که در روش تحلیل مورد کامل به دست آمده‌اند که کارایی بیشتر الگوریتم EM را نشان می‌دهد و با وجود اینکه هیچ قسمت از اطلاعات از مدل حذف نمی‌شوند، پارامترهایی با انحراف معیارهای پایین‌تر برآورد می‌شوند.

به‌طور کلی در هر پژوهش توجه به مقادیر گمشده و استفاده از روش‌های مناسب جانهی آنها امری مهم و ضروری است. چراکه حذف معمولی آنها موجب از دست رفتن بخشی از اطلاعات و کم شدن حجم نمونه شده و استفاده از روش‌های نامناسب نیز موجب برآوردهای نادقیق شده و در هر دو حالت، نتایج به دست آمده از اعتبار کافی برخوردار نخواهند بود. لذا مواقعی که به برخی از دلایل، اطلاعات ناقص است،

پژوهشگر باید با مقایسه روش‌های مختلف جانهی با شرایط مورد بررسی، مناسب‌ترین روش جانهی را برای به دست آوردن داده‌های کامل برگزیند.

منابع

آشفته، افشین (۱۳۹۲). بررسی روش‌های برخورد با داده‌های گمشده. *مجله اندیشه آماری*، ۲، ٤٠-٤٧.

افشاری صفوی، علیرضا؛ کاظم‌زاده قره‌چبق، حسین و رضایی، منصور (۱۳۹٤). مقایسه روش الگوریتم EM و روش‌های متداول جانهی داده‌های گمشده: مطالعه روی پرسشنامه خوددرمانی بیماران دیابتی، *مجله تخصص اپیدمیولوژی ایران*؛ ۱۱ (۳)، ٤۳ – ٥۱.

پورحسینقلی، محمدامین؛ علوی مجد، حمید؛ ابدی، علیرضا و پروانه‌وار، سیمین (۱۳۸٤). تحلیل درست‌نمایی ماکسیمم مدل رگرسیون لجستیک در حالتی که داده های متغیرهای پیشگو کامل نیستند ولی متغیرهای کمکی وجود دارند، *مجله اپیدمیولوژی ایران*، ۱ (۲)، ٦٥ – ۷۲.

رشیدی‌نژاد، آسیه و نواب‌پور، حمیدرضا (۱۳۸۹). مقایسه جانهی الگوریتم EM با دو روش جانهی میانگین و نمونه‌های جدید در آمارگیری‌های پانلی. *مجله بررسی های آمار رسمی ایران*، ۲۱ (۱)، ۸۹ – ۱۰۸.

زائری، فرید؛ اکبرزاده باغبان، علی‌رضا؛ کاظم‌زاده، مژگان؛ یاسری، مهدی و عباسی، علی‌محمد (۱۳۹۱). انواع گمشدگی در مطالعات طولی و روش‌های مبنی بر درست‌نمایی برای تحلیل آنها. *مجله علمی دانشگاه علوم پزشکی ایلام*، ٤، ۲۰۸ – ۲۲۲.

قاسمی، وحید (۱۳۸۹). *مدل‌سازی معادله ساختاری در پژوهش‌های اجتماعی با کاربرد Amos*. تهران: انتشارات جامعه‌شناسان.

Allison, P. D. (2005). *Imputation of categorical variables with PROC MI*. [accessed July 30, 2006]. http://www2.sas.com/proceedings/sugi30/113-30.pdf.

Arbuckle, J. L., & Wothke, W. (1999). *AMOS 4.0 user's guide* [Computer software manual]. Chicago: Smallwaters.

Bernaards, C. A.; Belin, T. R. & Schafer, J. L. (2007). Robustness of multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26, 1368–1382.

BMDP Statistical Software. (1992). *BMDP statistical software manual*. Los Angeles: University of California Press.

Bryk, A. S.; Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs.* Chicago: Scientific Software International.

De Leeuw, E. D.; Hox, J. J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, 19, 153–176.

Dempster, A. P.; Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39 (1), 1–22.

Doucleau, A. F.; Mauer, M.; Molenberghs, G. & Albert, A. (2015). Communications in Statistics – Simulation and Computation: A Simulation Study Comparing Multiple Imputation Methods for Incomplete Longitudinal Ordinal Data. *Communications in Statistics—Simulation and Computation*, 44, 1311-1338.

Fleiss, J. L.; Levin, B. & Paik, M. C. (2002). *Statistical Methods for Rates and Proportions*, 3rd ed. John Wiley & Sons.

Gellman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, New York.

Glynn, R. J. & Laird, N. M. (1983). *Regression Estimates and Missing Data: Complete Case Analysis*. Unpublished Manuscript, Department of Biostatistics, Harvard University.

Graham, J. W., & Hofer, S. M. (1991). *EMCOV.EXE users 'guide* [Computer software manual]. Unpublished manuscript, University of Southern California, Los Angeles.

Graham, J.; Hofer, S., & MacKinnon, D. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31,* 197–218. Doi: 10.1207/ s15327906mbr3102_3.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society: Series B, Methodological, 30,* 67–82.

Honaker, J.; King, G.; Blackwell, M. (2006). *Amelia software website*. Accessed December 15, 2006]. http://gking.harvard.edu/amela.

Honaker, J. & King, G. (2006). *What to do about missing values in time series cross-section data*. [Accessed December 17, 2006]. http://gking.harvard.edu/files/abs/pr-abs.shtml.

Horton, N. J.; Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57 (4), 229–232.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1), 79–90.

Imai, K.; King, G. & Lau, O. (2006). *Zelig software website*. [Accessed December15, 2006].http://gking.harvard.edu/zelig.

Insightful (2001). *S-PLUS (Version 6)* [Computer software]. Seattle, WA: Insightful.

Jo¨reskog, K. G., & So¨rbom, D. (2001). *LISREL (Version8.5)* [Computer software]. Chicago: Scientific Software International.

Kim, J. (2004). Finite sample properties of multiple imputation estimators. *Annals of Statistics, 32*, 766–783. Doi: 10.1214/009053604000000175.

King, G; Honaker, J.; Joseph, A. & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, 95, 49–69.

Little, R. J. & Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley New York.

Littell, R. C.; Milliken, G. A.; Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons; New York.

Marwala, T. (2009). *Computational Intelligence for Missing Data Imputation, Estimation and Management: Knowledge Optimization Techniques*, South Africa: University of Witwatersrand IGI Global 2009 ISBN 978-1-60566-336-4.

McKnight, P.; McKnight, K.; Sidani, S., & Figueredo, A. (2007). *Missing data: A gentle introduction*. New York, NY: Guilford Press.

Multilevel Models Project (1996). *Multilevel modeling applications—A guide for users of MLn.* [Computer software manual]. London: University of London, Institute of Education.

Muthe´n, L. K., & Muthe´n, B. O. (1998). *Mplus user's guide* [Computer software manual]. Los Angeles: Muthe´n & Muthe´n.

Neale, M. C.; Boker, S. M.; Xie, G., & Maes, H. H. (1999). Mx: *Statistical modeling (5th Ed.) [Computer software]*. Richmond: Virginia Commonwealth University, Department of Psychiatry.

Nirelli, L. M.; Larsen, M. D.; Croghan, I. T.; Schroeder, D. R.; Offord, K. P. & Hurt, R. D. (2005*) Comparison of methods for handling missing data in a collegiate survey of tobacco use Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.

Peng, C.; Harwell, M.; Liou, S., & Ehman, L. (2006). Advances in missing data methods and implications for educational research. In S. S. Sawilowsky (Ed.), *Real data analysis* (pp. 31–78). Charlotte, NC: New Information Age.

Peugh, J., & Enders, C. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525–556. Doi: 10.3102/00346543074004525.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987.

Robins, J. M., & Rotnitzky, A. (1992). *Recovery of information and adjustment for dependent censoring using surrogate markers*. Boston: Birkhauser.

Rubin, D. B. (1996). Multiple Imputation after 18+ Years (with discussion), *J. A. Stat. Asso,* 19, 473-489.

Salkind, N., & Rasmussen, K. (2007). *Encyclopedia of measurement and statistics.* Thousand Oaks, CA: Sage.

Stata. (2001). *Stata user's guide* [Computer software manual]. College Station, TX: Author.

Schafer, J. L. (1997a). *Analysis of incomplete multivariate data*, Chapman & Hall, New York.

Schafer, J. L. (1997b). *Introduction to multiple imputations for missing data problems*, viewed 6 May 2002, <www.stat.psu.edu/~jls/asa97/slide7.html>.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Book number 72 in the Chapman & Hall series Monographs on Statistics and Applied Probability. London.

Schimert, J.; Schafer, J. L.; Westerberg, T.; Fraley, C., & Clarkson, D. (2001). *Analyzing missing values in SPLUS*. Seattle, WA: Insightful.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association* 82, 528–550.

Templ, M. & Filzmoser, P. (2008). *Visualization of missing values using the R-package VIM*, Reserach report cs-2008-1, Department of Statistics and Probability Theory, Vienna University of Technology.

Templ, M; Kowarik, A. & Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods, *Computational Statistics & Data Analysis*, 55, 2793-2806.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL.

Von, Hippel P. (2004). Biases in SPSS 12.0 missing value analysis. *The American Statistician*, 58 (2), 160–164.

Wayman, J. C. (2003). *Multiple imputation for missing data: What is it and how can I use it*, in Annual Meeting of the American Educational Research Association, Chicago, IL, pp. 2- 16.

Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference* (Paper No. 267). Cary, NC: SAS Institute.

Young, W.; Weckman, G., & Holland, W. (2011). A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits. *Theoretical Issues in Ergonomics Science*, *12*, 15 – 43.