پژوهش‌های نوین در تصمیم‌گیری
دوره ۶، شماره ۱، بهار ۱۴۰۰، صص ۲۰۱-۲۲۹
نوع مقاله: پژوهشی

تحلیل پوششی داده‌ها با داده‌های گمشده

بهمن فصیحی[۱]، حسین عزیزی[۲*]، زینب قلیزاده گزور[۳]

۱. دانشجوی دکتری، گروه ریاضی، واحد پارس‌آباد مغان، دانشگاه آزاد اسلامی، پارس‌آباد مغان، ایران
۲. استادیار، گروه ریاضی، واحد پارس‌آباد مغان، دانشگاه آزاد اسلامی، پارس‌آباد مغان، ایران
۳. کارشناسی ارشد، گروه ریاضی، واحد پارس‌آباد مغان، دانشگاه آزاد اسلامی، پارس‌آباد مغان، ایران

تاریخ ارسال: ۱۳۹۹/۱۰/۲۱ تاریخ پذیرش: ۱۳۹۹/۱۲/۲۱

چکیده

داده‌های گمشده در کاربردهای تحلیل پوششی داده‌ها یک بیماری مزمن محسوب می‌شوند. خیلی از اوقات، متغیرهای مهم ورودی یا خروجی پوشش ناکامل دارند و یا اینکه واحدهای تصمیم‌گیری همه آمارهای لازم را گزارش نمی‌کنند. بنابراین، مقادیر گمشده در ورودی‌ها و خروجی‌ها را نمی‌توان با مدل‌های اصلی تحلیل پوششی داده‌ها مورد بررسی قرار داد. در این مقاله، روش‌هایی را برای پیدا کردن داده‌های گمشده در حالتی که داده‌ها قطعی هستند، ارائه می‌کنیم. در این مقاله پس از تشریح مفاهیم ضروری داده‌های گمشده، برخی از روش‌های جانهی داده‌های گمشده که موجب کاهش پیچیدگی تحلیل داده‌ها می‌شود، تشریح می‌شود. روش‌های مختلفی برای جانهی داده‌های بی‌پاسخ موجود است؛ از جمله روش‌های گوناگون جانهی ساده و جانهی چندگانه. این مقاله نخستین کوشش سیستماتیک برای بهره‌مندی از داده‌های حاوی مقادیر گمشده با بهره‌مندی از رویکردهای آماری در DEA است. به طور خاص، بررسی می‌کنیم که اگر درایه‌های خالی را در مجموعه داده‌ها نگه داریم و یک مقدار عددی خاص به آن‌ها اختصاص دهیم، چه اتفاقی می‌افتد. برای نشان دادن طرز کار روش‌های پیشنهادی، از این روش‌ها برای ارزیابی مجموعه‌ای از مدارس متوسطه دولتی یونان که برخی دارای مقادیر گمشده در ورودی یا خروجی هستند، بهره‌مندی خواهد شد.

کلیدواژه‌ها: تحلیل پوششی داده‌ها؛ اندازه‌گیری کارایی؛ داده‌های گمشده.

---

* نویسندهٔ مسؤول مقاله E-mail: azizhossein@gmail.com

۱- مقدمه

بی‌پاسخی در سرشماری‌ها و آمارگیری‌ها، امری است که هیچ پژوهشگر جوامع انسانی نمی‌تواند به‌طور کامل از آن پرهیز نماید. روش‌های بسیاری برای کاستن بی‌پاسخی‌های واحد آماری و قلم آماری معرفی و بررسی شده‌اند. به هر حال در برخی از مطالعات هنوز نرخ بی‌پاسخی‌ها بالاست و روش‌های تحلیلی که با وجود بی‌پاسخی بتوانند نتایج آماری معتبری در اختیار پژوهشگر قرار دهند، مورد نیاز است.

در یک سرشماری از جامعه، کوشش بر گردآوری اطلاعات از هر واحد در جامعه است. برای نمونه در یک سرشماری از یک جامعه انسانی با هر فرد در آن جامعه تماس گرفته می‌شود و سن، جنس، میزان تحصیلات و سایر ویژگی‌های آن فرد ثبت می‌شود. در یک نمونه‌گیری از جامعه، اطلاعات مشابه با سرشماری اما تنها برای برخی از افراد جامعه که در نمونه حضور دارند، ثبت می‌شود. در بسیاری از سرشماری‌ها و نمونه‌گیری‌ها، برخی از واحدها به تعدادی از پرسش‌ها پاسخ نمی‌دهند. بی‌پاسخی امری است که هیچ پژوهشگر جامعه انسانی نمی‌تواند به بخشی از نمونه‌های برنامه‌ریزی شده دسترسی پیدا کند. در هر آمارگیری، تماس با برخی افراد، خانوار، بنگاه‌ها و سایر انواع دیگر واحدهای نمونه‌گیری میسر نمی‌شود. این گونه عدم پاسخ‌ها چه در آمارگیری‌ها و چه در سرشماری‌ها در عمل برای واحدهای آماری مورد نظر به‌صورت فردی، خانوار یا تجاری مشترک می‌باشند.

بی‌پاسخی زمانی ایجاد می‌شود که یک واحد از واحدهای جامعه آماری به همه یا بخشی از پرسش‌های پرسشنامه آمارگیری پاسخ ندهند و این خود یکی از منابع بالقوه در ایجاد خطا در آمارگیری است.

روش‌های آماری استاندارد، عمدتاً برای داده‌های کامل تهیه و ارائه شده‌اند. اغلب داده‌های کامل به‌صورت یک ماتریس تصور می‌شوند که سطرهای آن مربوط به واحدهای آماری است که بر حسب موضوع، نام‌های مختلفی بر آن نهاده شده است و ستون‌های این ماتریس مربوط به متغیرهایی است که بر هر واحد اندازه‌گیری می‌شود. در ادبیات آماری این ماتریس را «ماتریس داده‌ها» نامگذاری کرده‌اند و فایل‌های به‌دست آمده به این طریق یک «فایل داده» می‌نامند. درایه‌های این ماتریس که تقریباً همیشه اعداد حقیقی هستند، مقادیری را برای متغیرهای پیوسته از قبیل سن، درآمد یا وزن ارائه می‌دهند یا پاسخ رسته‌ای هستند که ممکن است ترتیبی (مثل سطح سواد) یا اسمی (مثل نژاد و جنیست) باشند. در اینجا ماتریس داده‌های مورد بحث، برای برخی درایه‌ها مشاهده‌ای وجود ندارد. شاید معمولی‌ترین روش

کار با داده‌های ناقص، کنارگذاری همه واحدهایی که دارای عناصر داده‌ای گمشده یا ناقص در یکی از متغیرهایشان هستند، باشد. گرچه این روش، متداول بوده و بهره‌مندی فوری از روش‌های آماری مربوط به داده‌های کامل را میسر می‌سازد، اما دلایلی وجود دارند که این روش را نمی‌توان عموماً شیوه‌ای خوب به حساب آورد. یکم، کنارگذاری همه واحدهای با داده‌های گمشده، با کاهش حجم نمونه، موجب کاهش توان و کارایی نتایج بسیاری از تحلیل‌ها می‌شود. دوم، اگر واحدهایی که کنارگذاری می‌شوند مشخصه یا مشخصه‌هایی گوناگون از مشخصه‌های واحدهایی که باقی می‌مانند داشته باشند، کنارگذاری واحدهای ناقص، تورشی زیاد برآوردهای پایانی را موجب می‌شود. سوم، اگر الگوی نمونه‌گیری مثلاً نمونه‌گیری با احتمال متناسب با اندازه باشد، برآوردهای حاصل در حالت کلی به دلیل به هم خوردن نظام وزن‌ها تورش می‌شوند.

تحلیل پوششی داده‌ها (DEA)[1] یک رویکرد برنامه‌ریزی ریاضی ناپارامتری با کاربرد بسیار گسترده برای تحلیل بهره‌وری و کارایی است [۱]. اصل پایه‌ای در DEA این است که به‌جای اینکه داده‌ها را وادار کنیم که پیرو یک شکل تابعی معین شده دلخواه و بدون انعطاف باشند، داده‌ها را به همان صورت که هستند، مورد بهره‌مندی قرار دهیم. در نتیجه، مفید بودن DEA مستقیماً بستگی به کیفیت داده‌ها و نیز کمیت داده‌ها دارد. این مطلب به خوبی روشن شده است که نتایج DEA نسبت به خطاهای داده‌ها، برون‌هشته‌ها و مشکلات دیگر مربوط به کیفیت داده‌ها حساس است. هم‌اکنون پژوهش‌های زیادی برای بهبود مقاومت DEA برای کار با داده‌های با کیفیت پایین‌تر در جریان است (برای بحث و منابع دیگر، به‌عنوان نمونه، رک. پست[2] و همکاران [۲]، کوسمانن[3] و همکاران [۳]). همچنین، این مطلب روشن شده است که DEA به خاطر ماهیت ناپارامتری و چندبعدی خود، عموماً به تعداد زیادی مشاهدات نیاز دارد تا نتایج آن از نظر آماری معنی‌دار باشد (برای بحث درباره خواص آماری برآورد کننده‌های کارایی DEA، به‌عنوان نمونه، رک. سیمار[4] و ویلسون[5] [۴]).

متغیرهای ورودی-خروجی بالقوه مهم غالباً پوشش ناکافی دارند و یا اینکه بنگاه‌ها همه آمارهای لازم را گزارش نمی‌کنند. مسئله *مشاهدات گمشده* یک بیماری مزمن در کاربردهای DEA است، که هم بر کیفیت و هم بر کمیت داده‌ها تأثیر می‌گذارد. البته، راه درمان فوری این است که وقت و کوشش بیشتری را برای گردآوری داده‌ها صرف کنیم. با این حال، بر خلاف علوم طبیعی که در آن داده‌ها در تجربیات آزمایشگاهی یا کارآزمایی‌های میدانی ایجاد

می‌شوند، کاربردهای DEA غالباً متکی بر داده‌های مشاهده‌ای غیرتجربی هستند. به گفته گریلیچز[4] [5]: «اکثر کارهای ما بر روی داده‌های ‹پیدا شده› است، یعنی داده‌هایی که به وسیله فرد دیگری گردآوری شده‌اند و چه بسا که هدف گردآوری آن دلیل کاملاً گوناگونی بوده است (ص. ۱۴۶۶).» در این شرایط، ما غالباً مجبوریم داده‌های گمشده را به‌عنوان یک ویژگی نامطلوب داده‌های دنیای واقعی بپذیریم. امروزه، مدل‌های رگرسیون داده‌های پانلی به طور متداول از داده‌های نامتوازن که حاوی درایه‌های خالی هستند، بهره‌مندی می‌کنند. بر عکس، کار با داده‌های گمشده تنها با تذکرات گذرایی در ادبیات DEA همراه بوده است (استثنای قابل توجه آن کائو[7] و لئو[8] [۶] هستند، که از مجموعه‌های فازی برای مدل‌سازی دامنه‌های صحیح برای داده‌های گمشده بهره‌مندی کرده‌اند). با این حال، مشاهدات گمشده در محیط DEA نیز به اندازه تحلیل رگرسیون شایع است. لذا، کنارگذاری دلخواه داده‌های گمشده از ماتریس‌های داده‌ها همچنان رویکرد استاندارد برای کار با این مسئله به شمار می‌رود. برخی از نویسندگان به طور شفاف خارج کردن بنگاه‌ها و/یا متغیر را از مجموعه داده‌ها اعلام می‌کنند، ولی به طور عقلانی می‌توان تصور کرد که خیلی دیگر از نویسندگان مشاهدات را در مرحله قبل از پردازش از داده‌ها کنارگذاری می‌کنند، بدون اینکه در معیارهای انتخاب به طور صریح به این امر اشاره کنند. مشکل این است که معمولاً چندین راه برای درست کردن داده‌های متوازن وجود دارد و انتخاب اینکه چه بنگاه‌ها و متغیرهایی در مجموعه داده‌ها دخالت داده شوند، می‌تواند بر نتایج تأثیر بگذارد.

گاردیجان[9] و لوکاچ[10] [۷] کارایی نسبی صنایع غذایی و آشامیدنی در کشورهای منتخب اتحادیه اروپا را با بهره‌مندی از DEA با داده‌های گمشده اندازه‌گیری کردند. چن[11] و همکاران [۸] یک ارزیابی چندمعیاره با بهره‌مندی از DEA مقادیر صحیح با داده‌های گمشده انجام دادند. دیورت[12] و همکاران [۹] برای کار با اطلاعات گمشده در DEA، از رویکرد ماتریس رتبه پایین برای پیش پردازش داده‌های گمشده استفاده کردند. استید[13] و ویت[14] [۱۰] با بهره‌مندی از روش‌های جانهی داده‌های گمشده چندگانه در تحلیل مرز تصادفی، داده‌های بزرگراه محلی انگلیسی را تشریح کردند.

در این مقاله پس از تشریح مفاهیم ضروری داده‌های گمشده، برخی از روش‌های جانهی داده‌های گمشده که موجب کاهش پیچیدگی تحلیل داده‌ها می‌شود، تشریح می‌شود. روش‌های مختلفی برای جانهی داده‌های بی‌پاسخ موجود است؛ از جمله روش‌های گوناگون جانهی ساده

و جانهی چندگانه که در این مقاله تنها برخی از روش‌های جانهی ساده بیان خواهد شد. این مقاله نخستین کوشش سیستماتیک برای بهره‌مندی از داده‌های حاوی مقادیر گمشده با بهره‌مندی از رویکردهای آماری در DEA است. به طور خاص، بررسی می‌کنیم که اگر درایه‌های خالی را در مجموعه داده‌ها نگه داریم و یک مقدار عددی خاص به آن‌ها اختصاص دهیم، چه اتفاقی می‌افتد.

ادامه مقاله به‌صورت زیر سازماندهی شده است. در بخش ۲ سازوکار داده‌های گمشده بیان می‌شود. بخش ۳ روش‌های کار در مورد داده‌های گمشده ارائه می‌کند. بخش ۴ مدل BCC را معرفی می‌کند. برای نشان دادن فایده عملی این نتایج، در بخش ۵ (به‌عنوان یک بررسی موردی)، کاربرد عملی آن را در ارزیابی مجموعه‌ای از مدارس متوسطه دولتی یونان بررسی می‌کنیم. بخش ۶ ملاحظات پایانی را ارائه می‌کند.

۲- سازوکار داده‌های گمشده

در حالت کلی وقتی می‌توانیم استنباط‌های معتبر در مورد داده‌های ناکامل داشته باشیم که سازوکار داده‌های گمشده به طور کامل شناخته شده باشند. ولی از آنجایی که این سازوکار تحت کنترل پژوهشگر نیست، شناخت دقیق آن امکان‌پذیر نیست. بنابراین، در هر بررسی فرضیاتی درباره این سازوکار پذیرفته می‌شود. این فرضیات در واقع در پاسخ به این سؤال که «چرا گمشدگی رخ داده و یا به طورخاص‌تر، آیا مقادیر گمشده ارتباطی با سوال‌های کاربردی پژوهش دارند یا نه؟» شکل می‌گیرند. چنانچه دلایل گمشدگی با پیامد مورد نظر مرتبط نباشد، مشکل خاصی در تحلیل داده‌ها وجود نخواهد داشت. ولی در صورتی که این ارتباط وجود داشته باشد، ممکن است منجر به برآوردهای تورش شود؛ زیرا داده‌های در دسترس با داده‌های کامل گوناگون خواهند بود. مسلماً اعتبار تحلیل‌ها به برقراری یا عدم برقراری این فرضیات وابسته است. لیتل[۵] و روبین[۴] [۱۱] سازوکارهای گمشدگی را به سه دسته تقسیم‌بندی کردند:

۱- **گمشدگی کاملاً تصادفی (MCAR)**[۷]: این نوع گمشدگی زمانی اتفاق می‌افتد که احتمال مشاهده نشدن پاسخ در یک زمان به هیچ یک از پاسخ‌های مشاهده شده و مشاهده نشده بستگی نداشته باشد. یعنی اگر نشانگر r_i را به‌صورت زیر تعریف کنیم:

$$r_i = \begin{cases} 0, & y_i \text{ is observed} \\ 1, & \text{otherwise} \end{cases} \qquad (1)$$

آنگاه

$$p(r_i \mid y_i^{\text{observed}}, y_i^{\text{missing}}, X_i) = p(r_i \mid X_i) \qquad (2)$$

به‌عنوان نمونه فرض کنید افراد مورد بررسی در مناطق مختلف شهری ساکن هستند. چنانچه دلیل عدم ثبت داده‌های افراد، مسایلی مانند بارش برف و یا ترافیک سنگین در روز تعیین شده برای رسیدن به مرکز آزمایشات باشد، گمشدگی کاملاً تصادفی رخ داده است؛ زیرا در این حالت گمشدگی داده‌ها ارتباطی به مشاهدات پیشین ندارد.

۲- **گمشدگی تصادفی (MAR)**[۸]: گمشدگی تصادفی زمانی اتفاق می‌افتد که احتمال گمشدگی تنها به پاسخ‌های مشاهده شده وابسته باشد ولی به پاسخ‌های مشاهده نشده بستگی نداشته باشد، یعنی:

$$p(r_i \mid y_i^{\text{obs}}, y_i^{\text{miss}}, X_i) = p(r_i \mid y_i^{\text{obs}}, X_i) \qquad (3)$$

همانطور که مشاهده می‌شود در این نوع گمشدگی، احتمال گمشدن ممکن است به مشاهدات پیشین وابسته باشد. این سازوکار نسبت به MCAR معقول‌تر به نظر می‌رسد، هر چند باعث ایجاد محدودیت‌های بیش‌تر در تحلیل خواهد شد. به‌عنوان نمونه ممکن است بررسی به گونه‌ای طراحی شود که در صورتی که مقدار متغیر پاسخ برای فردی خارج از محدوده کلینیکی خاصی ثبت شود، آنگاه آن فرد از بررسی خارج شود. در این صورت گمشدگی در پاسخ تحت کنترل پژوهشگر است و تنها به مؤلفه‌های مشاهده شده بستگی دارد.

۳- **گمشدگی غیرتصادفی (MNAR)**[۹]: این گمشدگی زمانی اتفاق می‌افتد که احتمال گمشدگی دست‌کم به یکی از پاسخ‌هایی که باید مشاهده می‌شد ولی گمشده است، وابسته باشد؛ یعنی $p(r_i \mid y_i^{\text{obs}}, y_i^{\text{miss}}, X_i)$ دست‌کم به یکی از مولفه‌های y_i^{miss} وابسته است. برای روشن‌تر شدن این نوع از سازوکار گمشدگی به دو نمونه زیر توجه کنید:

I. فرض کنید هدف اندازه‌گیری کیفیت زندگی افراد باشد، چنانچه افرادی از سطح کیفیت زندگی بالا یا پایینی برخوردار باشند، از پر کردن پرسشنامه و کامل کردن بررسی آن اجتناب کنند، گمشدگی غیرتصادفی رخ داده است.

II. بررسی روی بیماران تنفسی در شهری خاص را در نظر بگیرید، اگر علت گمشدگی در

داده‌های این بررسی مهاجرت بیماران باشد و چنانچه علت مهاجرت گسترش مشکل تنفسی افراد به دلیل آلودگی هوای آن شهر باشد، گمشدگی MNAR رخ داده است چرا که سازوکار گمشدن وابسته به مقادیرگمشده است.

چنانچه سازوکار گمشدگی MCAR و MAR باشد، گمشدگی از نوع چشم‌پوشی پذیر است و در مقابل گمشدگی غیرتصادفی، MNAR، گمشدگی چشم‌پوشی ناپذیر یا آگاهی‌بخش نامیده می‌شود. امروزه بسیاری از پژوهشگران روش‌هایی را برای مقابله با داده‌های گمشده ارائه کرده‌اند که بیشتر آن‌ها وابسته به نوع سازوکار گمشدگی داده‌هاست. پس در ابتدا باید نوع این سازوکار مشخص باشد تا بتوان روش تحلیلی مناسبی را انتخاب کرد. هیچ گاه به طور قاطع نمی‌توان درباره احتمال گمشدگی نظر داد زیرا معمولاً به طور دقیق از مقادیر گمشدگی مطلع نیستیم. سازوکار گمشدگی MCAR قابل شناسایی از طریق t-test معروف به آزمون لیتل است. می‌توان گمشدگی MAR را نیز به روش فلیس و همکاران [۱۲] شناسایی کرد. ولی در حقیقت در بسیاری از داده‌ها MNAR برقرار است و باید توجه داشت که هیچ تجزیه و تحلیلی از داده‌های دارای گمشدگی چشم‌پوشی ناپذیر بدون انجام آزمون حساسیت کامل نیست (ابراهیم و همکاران [۱۳] و کریم‌لو و همکاران [۱۴]).

باید توجه داشت که بین الگوی گمشدگی و سازوکار گمشدگی تفاوت وجود دارد. الگوی گمشدگی نشان‌دهنده ترکیب داده‌های گمشده و مشاهده شده در مجموعه داده‌هاست. در حالی که سازوکار گمشدگی رابطه ممکن بین داده‌های گمشده و مشاهده شده را توضیح می‌دهد.

به طور کلی می‌توان الگوهای گمشدگی را به دو دسته تقسیم کرد:

- **الگوی یکنواخت**: این الگو به الگوی انصراف نیز مشهور است و به مواردی اطلاق می‌شود که در یک زمان خاص از بررسی خارج می‌شوند و دیگر باز نمی‌گردند (یعنی از یک جایی به بعد پاسخ وجود ندارد).
- **الگوی متناوب**: داده‌های مربوط به متغیر پاسخ در این الگو از گمشدگی به طور غیریکنواخت ثبت شده است.

به علاوه الگوهای دیگری از گمشدگی نیز مانند: *الگوی تک متغیره*، *الگوی بی‌پاسخی واحد*، *الگوی متغیر پنهان* و *الگوی گمشدگی* با برنامه نیز وجود دارد.

از دیدگاه عملی امروزه دیگر تمایزی بین الگوی گمشدگی داده‌ها وجود ندارد. زیرا

روش‌های برآورد مانند MLE و روش جانهی چندگانه برای همه الگوهای گمشدگی از داده‌ها مناسب هستند.

از آنجایی که هیچگاه نرخ بی‌پاسخی علی‌رغم همه کوشش‌ها به صفر نخواهد رسید، باید با به‌کارگیری روش‌های علمی مناسب نسبت به جایگزینی یا جانهی داده‌های بی‌پاسخی اقدام نمود.

از عواقبی که با وجود بی‌پاسخی (داده‌های گمشده)، ممکن است در تحلیل‌های آماری رخ دهد، می‌توان موارد ذیل را برشمرد:

۱- کاهش کارایی استنباط‌های آماری؛

۲- پیچیدگی در پردازش و تحلیل داده‌ها؛

۳- ایجاد تورشی ناشی از اختلاف بین داده‌های مشاهده نشده و داده‌های مشاهده شده.

با توجه به اثرات سوء بی‌پاسخی در برآوردهای یک آمارگیری و عدم امکان پرهیز از آن، یکی از راه‌های مقابله با این مشکل، جانهی داده‌های گمشده است. هدف از جانهی را می‌توان در دو مورد زیر خلاصه کرد [۱۵]:

۱- دادن امکان به کاربران پایانی داده‌ها با دستورالعمل و خروجی‌های استاندارد در بهره‌مندی از ابزارهای تجزیه و تحلیل موجود برای هر مجموعه‌ای که شامل داده‌های گمشده نیز می‌باشند.

۲- ارائه استنباط‌های آماری معتبر برای داده‌های شامل بی‌پاسخی از طریق به‌کارگیری روش‌های جانهی.

در طی سال‌های متمادی روش‌های متعددی برای جانهی مقادیر گمشده پیشنهاد شده‌اند و نرم‌افزارهای مربوط تهیه شده و توسعه پیدا کرده‌اند. هر یک از این روش‌ها در شرایط خاص خود خوب عمل می‌کنند. روش‌های مورد استقاده ضمن اینکه تابع عوامل مختلفی است به نوع بی‌پاسخی نیز مربوط است.

معمولاً بی‌پاسخی در آمارگیری‌ها از دو نوع‌اند: (۱) بی‌پاسخی واحد (واحد بی‌پاسخ)، (۲) بی‌پاسخی پرسش (پاسخ یا قلم بی‌پاسخ). وقتی بی‌پاسخی شامل واحد باشد بدان مفهوم است که همه پرسش‌های مربوط به آن واحد فاقد مشاهده یا فاقد اعتبار است (که این می‌تواند حتی شامل مواردی باشد که در مرحله ویرایش پذیرفته نمی‌شوند). اگر برای یک واحد آماری تنها اطلاعات طرح آمارگیری از قبیل موقعیت مکانی خانوار یا مشخصاتی که در عملیات فهرست‌برداری قبل از آمارگیری گردآوری شده است در اختیار باشد، این واحد آماری نیز بی‌پاسخ تلقی می‌شود [۱۶] (ص. ۵). در بی‌پاسخی نوع دوم، جواب یک یا چند پرسش توسط

پاسخگو ارائه نمی‌شود (که این هم می‌تواند شامل مواردی باشد که در مرحله ویرایش پذیرفته نمی‌شوند). اما سایر پرسش‌های پاسخ داده شده در حدی قابل توجه است که پرسش‌های پاسخ داده شده آن واحد مورد نظر قرار می‌گیرد. در اینجا فقط جانهی بی‌پاسخی پرسش مطرح خواهد شد.

نحوه بی‌پاسخی پرسش دو گونه است:

۱- **جانهی ساده**: در جانهی ساده برای هر پرسش بی‌پاسخ تنها یک مقدار جایگزینی صورت می‌گیرد.

۲- **جانهی چندگانه**: در جانهی چندگانه برای هر پرسش بی‌پاسخ چندین مقدار جایگزینی صورت می‌گیرد.

روش‌های جانهی بی‌پاسخی پرسش از دید کلی به دو دسته تقسیم می‌شوند:

۱- **جانهی تصادفی** (**Stochastic**): که در آن برای هر یک از موارد بی‌پاسخی مقادیری تصادفی از روی مقادیر مشاهده شده یا از یک توزیع پیش‌بینی شده جانهی می‌شوند.

۲- **جانهی قطعی**: که در آن هر بار (در جانهی ساده یک بار و در جانهی چندگانه چند بار) تنها یک مقدار ثابت برای پرسش بی‌پاسخ جانهی می‌شود.

۳- روش‌های کار با داده‌های گمشده

۱-۳- جانهی با میانگین

گرچه این روش ساده‌ترین نوع جانهی است اما شاید بتوان گفت غیرجذاب‌ترین نوع جانهی نیز می‌باشد. جانهی همه مقادیر بر پایه میانگین موجب اختلال در توزیع داده‌ها و کم برآورد شدن واریانس می‌شود.

در این روش میانگین کل مشاهدات هر متغیر و یا میانگین‌های سلولی جایگزین مقدار برای پرسش‌های بی‌پاسخ می‌شود. این مورد در مواردی به کار می‌رود که متغیر مورد نظر کمی باشد و در نتیجه میانگین برای آن مفهوم داشته باشد، مانند درآمد، هزینه و سن.

اگر چه جانهی با میانگین روش نسبتاً آسانی است و علاوه بر آن سبب کاهش تورشی برآورد پارامترهایی مانند میانگین و مجموع می‌شود، ولی دو اشکال مهم در آن دیده می‌شود. یکی این که شکل توزیع Y (متغیر مورد نظر) را در بین اعضای نمونه دگرگون می‌کند. دوم اینکه چون به‌جای تعدادی از اعضای نمونه (بی‌پاسخی‌ها) یک مقدار ثابت مانند میانگین جایگزین می‌شود، تغییرات در بین اعضای نمونه کاهش پیدا می‌کند و این سبب

می‌شود که واریانس حاصل (به میزانی که تقریباً با نرخ پاسخ برابر است) کمتر از واریانس واقعی نشان داده شود. در اینجا منظور از واریانس واقعی، واریانس همه اعضای نمونه است، به شرط آنکه بی‌پاسخی وجود نداشته باشد. بنابراین، جانهی با مقدار میانگین برای تحلیل‌های ساده‌ای کاربرد دارد که برآوردهای نقطه‌ای چون میانگین یا مجموع بدون توجه به واریانس مورد نظر باشند.

گاهی در طرح‌های آماری از روش جانهی با مقدار میانگین کل، بدون تشکیل سلول‌های جانهی بهره‌مندی می‌شود. یعنی این که میانگین کل پاسخ داده‌های نمونه، جایگزین تک تک پاسخ نداده‌ها می‌شود. در چنین حالتی اگرچه بهره‌مندی از «جانهی با مقدار میانگین کل» بسیار آسان است و کم دردسر دارد، اما ضمن آن که دو اشکال اشاره شده را به شکل شدیدتری دارد، سبب کاهش تورشی حاصل از بی‌پاسخی نیز نمی‌شود. لذا بدون تشکیل سلول‌های جانهی، بهره‌مندی از روش فوق توصیه نمی‌شود.

در روش جانهی میانگین سلول، ابتدا با بهره‌مندی از برخی متغیرهای کمکی سلول‌های جانهی تشکیل داده می‌شود و سپس میانگین مقادیر مشاهده شده آن سلول‌ها جایگزین موارد گمشده از همان سلول می‌شود. تشکیل سلول‌ها معمولاً بر پایه متغیرهای رده‌ای صورت می‌پذیرد. این روش در صورتی برای میانگین اندازه صفت یا کل آن برآورد نااریبی به‌دست می‌دهد که متغیر گمشده تنها به متغیرهای کمکی بستگی داشته باشد که به کمک آن‌ها سلول جانهی تشکیل می‌شود.

در به کارگیری روش جانهی با میانگین توصیه می‌شود که در محاسبه برآورد ماتریس واریانس-کوواریانس در مخرج کسر از $1/(n-m-1)$ به‌جای $1/(n-1)$ بهره‌مندی شود که در آن n اندازه نمونه اصلی و m تعداد مشاهدات گمشده برای هر زوج متغیر یا تنها یک متغیر در مقابل متغیر دیگر می‌باشد. به این روش راهبرد جانهی میانگین تصحیح شده نیز گفته می‌شود.

در این راستا پیشنهاد دیگری برای تصحیح واریانس وجود دارد و آن جانهی بیش از یک مقدار برای موارد گمشده یا بی‌پاسخی است [۱۷]. به‌عنوان نمونه به جای جایگزینی میانگین به جای همه مشاهدات گمشده، نصف مشاهدات گمشده با $\bar{y}_r+\sqrt{(n+r-1)/(r-1)}D_r$ و نصف دیگر با $\bar{y}_r-\sqrt{(n+r-1)/(r-1)}D_r$ جانهی شوند که در آن r تعداد موارد با مقدار پاسخ، $\bar{y}_r$ میانگین مقادیر بابت ارائه پاسخ و $D_r^2=\frac{1}{r}\sum_{i=1}^{r}(y_i-\bar{y}_r)^2$ است.

سونگ[۲۱] و شفرد[۲۲] [۱۸] از شرایط بهره‌مندی از روش کوهن[۲۳] [۱۷] و روش‌های دیگر جانهی در داده‌کاوی را ارائه کردند.

۲-۳- روش الگوریتم EM

این الگوریتم یک روش جانهی قطعی مدل مبنا است که از دو گام تشکیل می‌شود [۱۹]:

گام E: در این گام امید ریاضی لگاریتم درستنمایی بر پایه داده‌های مشاهده شده و برآورد جاری پارامترها محاسبه می‌شوند.

گام M: برآورد پارامترها از روش بیشینه درستنمایی بر مبنای مقادیر جاری آماره‌های بسنده[۲۴] بهنگام می‌شوند.

این الگوریتم سپس به روش تکرار ادامه پیدا می‌کند تا جایی که اختلاف بین برآورد پارامترها در دو گام متوالی بر پایه یک معیار از قبل تعیین شده به همگرایی برسند. آخرین گام E میانگین مقادیر هر داده گمشده بر پایه آخرین برآورد پارامترها را محاسبه کرده که به‌عنوان مقادیر جانهی مورد بهره‌مندی قرار می‌گیرند.

گرچه این الگوریتم را می‌توان برای جانهی هر یک از مقادیر گمشده به‌صورت انفرادی به کار برد اما غالباً به طور مستقیم برای برآورد پارامتر جامعه مورد بهره‌مندی قرار می‌گیرد. در صورتی که فرض نرمال بودن برای داده‌ها برقرار باشد، امید ریاضی آماره‌های بسنده در گام E و برآوردهای بیشینه درستنمایی در گام M را می‌توان به راحتی به دست آورد. اما برای سایر توزیع‌ها کار ممکن است به این سادگی نباشد. همگرایی این روش پایین و هیچ تضمینی برای این گام همگرایی مخصوصاً برای داده‌های با حجم کم وجود ندارد. مزیت الگوریتم EM پایدار بودن همگرایی آن است، بدین مفهوم که تکرار موجب افزایش درستنمایی می‌شود.

کتلیر[۲۵] و همکاران [۲۰] با انجام شبیه‌سازی اثر دو جانهی در کاهش تورشی برآورد فعالیت جسمی دختران دوره دبیرستان را بررسی کردند. آن‌ها به دو شیوه تصادفی و سیستماتیک داده گمشده ایجاد کردند و با بهره‌مندی از دو روش الگوریتم EM و جانهی چندگانه فعالیت جسمی را برآورد کردند و تورشی برآورد را در این دو جانهی با هم مقایسه کردند. در گمشدگی تصادفی برآورد در هر دو روش جانهی غیر تورش بود. در گمشدگی سیستماتیک، تورشی برآورد فعالیت جسمی دو روش جانهی تفاوت معناداری نداشتند.

۳-۳- جانهی رگرسیون پیش‌بینی شده

این روش یک روش جانهی قطعی مدل مبنا است که در آن مقادیر گمشده بر پایه پیش‌بینی از یک مدل رگرسیونی برای همه موارد جایگزین می‌شوند. مقدار پیش‌بینی شده $\hat{y}_i$ بهترین پیش‌بینی کننده i-امین مقدار مشاهد نشده y_i تحت شرایط مدل ابرجامعه زیر است به شرط آنکه این مدل برای مقادیر باپاسخ و بی‌پاسخ هر دو برقرار باشد:

$$E(Y_i)=\alpha+\beta x_i,\quad V(Y_i)=\sigma^2,\quad \mathrm{cov}(Y_i,Y_j)=0,\quad i\neq j \qquad (۴)$$

جانهی رگرسیونی پیش‌بینی شده را نیز می‌توان در داخل هر رده جانهی مور بهره‌مندی قرار داد.

حالت عدم پاسخ یک متغیره را در نظر بگیرید که در آن $Y_1,\ldots,Y_{k-1}$ به طور کامل مشاهده شده‌اند و برای r مشاهده نخست مشاهده شده و برای $n-r$ مشاهده آخر گمشده است. جانهی رگرسیونی، رگرسیون Y_k را روی $Y_1,\ldots,Y_{k-1}$ بر پایه r عضو کامل به دست می‌آورد و سپس به جای مقادیر گمشده از رگرسیون مقادیر پیش‌بینی شده قرار داده می‌شود. به خصوص فرض کنید برای مورد i، y_{ik} گمشده است و $Y_{i1},\ldots,Y_{i,k-1}$ مشاهده شده‌اند. مقدار گمشده با بهره‌مندی از معادله رگرسیونی زیر جانهی می‌شود:

$$\hat{y}_{ik}=\tilde{\beta}_{k.12\ldots k-1}+\sum_{j=1}^{n}\tilde{\beta}_{kj.12\ldots k-1}y_{ij} \qquad (۵)$$

که $\tilde{\beta}_{k.12..k-1}$ عرض از مبدأ و $\tilde{\beta}_{kj.12..k-1}$ ضریب Y_j در رگرسیون Y_k روی $Y_1,\ldots,Y_{k-1}$ بر پایه r مورد کامل است. اگر متغیر مشاهده شده برای یک متغیر رسته‌ای، متغیر کاذب باشد، پیش‌بینی (۵) میانگین پاسخ داده شده‌ها داخل رده‌های تعریف شده به وسیله این متغیر است. به طور کلی ممکن است رگرسیون برای بهبود پیش‌بینی شامل متغیرهای پیوسته و رسته‌ای، اثرات متقابل و شکل‌های پارامتری محدود شده باشد.

۳-۴- بهره‌مندی از آنتروپی در جانهی

اصل امساک را ویلیام اوکام این طور بیان می‌کند: «علل را نباید فراتر از ضرورت و بیش از حد لازم بیان کرد». تعبیر این اصل در آمار بیان می‌کند که تحت شرایط معین برای معرفی یک توزیع باید توزیعی معرفی شود که بیش از شرایط معین ذکر شده، حاوی اطلاعات بیشتر نباشد. برای این کار کافی است توزیعی از خانواده توزیع‌ها که در آن شرایط صدق می‌کنند،

انتخاب شود که آنتروپی آن بیشینه باشد. از این رو «اصل آنتروپی» به‌عنوان یک ملاک در انتخاب‌ها و ملاک‌های آماری مورد بهره‌مندی قرار گرفت.

فرض کنید $x_1,\ldots,x_{n+1}$ داده‌های آماری از $n+1$ مشاهده نامنفی باشند که مقدار $x_{n+1}=x$ معلوم نیست. می‌بایستی x چنان تعیین شود که آنتروپی توزیع حاصل از مقادیر $x_1,\ldots,x_n$ و x بیشینه باشد. به کمک مقادیر نامنفی $x_1,\ldots,x_n,x$، متغیر تصادفی X را با مقادیر $x_1,\ldots,x_n,x$ و احتمال‌های زیر در نظر می‌گیریم:

$$p_i=P(X=x_i)=\frac{x_i}{x+\sum_{i=1}^{n}x_i},\quad i=1,\ldots,n$$
$$P(X=x)=\frac{x_i}{\sum_{i=1}^{n}x_i} \qquad (۶)$$

احتمال‌های (۶) به وضوح یک توزیع احتمال روی مقادیر داده‌های $x_1,\ldots,x_n$ و مقدار نامعلوم x تعریف می‌کنند. حال x را چنان تعیین می‌کنیم که آنتروپی این توزیع بیشینه باشد.

قضیه ۱: مقداری از x که توزیع معرفی شده در (۶) را بیشینه می‌کند به‌صورت زیر است:

$$x=\sqrt[s]{x_1^{x_1}\times x_2^{x_2}\times\cdots\times x_n^{x_n}} \qquad (۷)$$

که در آن $s=\sum_{i=1}^{n}x_i$.

قضیه ۲: فرض کنید از $n+2$ مقدار نامنفی $x_1,\ldots,x_n,x,y$ دو مقدار x و y نامعلوم باشند. مقادیری از x و y که آنتروپی توزیع X با مقادیر $x_1,\ldots,x_n,x,y$ با احتمال‌های

$$p_i=\frac{x_i}{s+x+y},$$
$$p_x=\frac{x}{s+x+y}, \qquad (۸)$$
$$p_y=\frac{y}{s+x+y}$$

را بیشینه می‌کند عبارت است از:

$$x = y = \sqrt[s]{x_1^{x_1} \times x_2^{x_2} \times \cdots \times x_n^{x_n}} \qquad (۹)$$

در این روابط $s = \sum_{i=1}^{n} x_i$.

نتیجه ۱: جانهی x به وسیله بیشینه آنتروپی واریانس را نسبت به جانهی x بـه وسـیله میانگین حسابی افزایش می‌دهد.

تذکر: در برآورد داده‌های گمشده اگر دو مقدار گمشده داشته باشیم چه این دو مقـدار را هم زمان و چه به این صورت که اول یکی را بـرآورد کـرده و از آن بانضـمام سـایر داده‌هـا برای برآورد داده گمشده بعدی بهره‌مندی کنیم نتیجه یکسان است.

۵-۳- روش داده افزایی در جانهی

داده افزایی یک روش بیز تکراری است کـه توسـط تـانر[۶] و ونـگ[۷] [۲۱] پیشـنهاد شـد. در ایـن روش دو توزیع در نظر گرفته می‌شود: یکی توزیع داده‌ها و دیگری توزیع پیشـین پارامترهـا. مشابه الگوریتم EM، این روش نیز شامل دو مرحله است: یکی مرحله اول یا مرحله I، کـه در آن مقادیر گمشـده بـا بهره‌منـدی از توزیـع پیش‌بینـی شـده بـرای داده‌هـا بـا بـه بکـارگیری برآوردهای جاری پارامترها جانهی می‌شوند. مرحله دوم یا مرحله P، که پارامترهـا بـر پایـه توزیع پیشین آن‌هـا و بـه کـارگیری مقـادیر مشـاهده شـده و مقـادیر جـانهی شـده بـرآورد می‌شوند. روش آن‌ها در ساختن مجموعه‌های داده‌های کامل، خیلی بـا نمونـه گیـبس[۸] مـرتبط است. در این روش از روابط بین متغیرها برای به دست آوردن مقادیر جانهی بـه طـور کـارا بهره‌مندی می‌شود. در صورتی که توزیع فرضی تقریباً با داده‌ها در توافق باشـد، ایـن روش معمولاً برآوردهای خوبی نیز برای واریانس به دست می‌دهد. تنها عیب این روش، نیاز آن بـه تکرار است که ممکن است مشابه الگوریتم EM همگرایی آن کند باشد.

۶-۳- جانهی بی‌درنگ قطعی

این روش به دلیل سادگی و دارا بودن معنا و مفهوم برای افرادی کـه در کارهـای آمـارگیری مشارکت دارند اما زمینه آماری قوی ندارند، یکی از معروفترین انواع روش جـانهی اسـت کـه در آن از هیچ مدل آماری صریحی بهره‌مندی نمی‌شود. عیـب عمـده ایـن روش آن اسـت کـه

نمی‌تواند مقادیر مشخصه برای افرادی را پوشش دهد که دارای خصیصه‌ها معینی بوده و هیچ یک از افراد دارنده این خصیصه به پرسش مورد نظر پاسخ نداده باشند. در این قبیل جانهی روش‌های متعددی به کار گرفته می‌شوند که در زیر معروف‌ترین آن‌ها شرح داده می‌شود.

فرض کنید یک نمونه nتایی از N واحد انتخاب می‌کنیم و r تا از n مقدار نمونه‌گیری شده یک بردار متغیرها $Y=(Y^{\text{obs}},Y^{\text{miss}})$ ثبت شده‌اند که Y^{obs} و Y^{miss} به ترتیب مربوط به بخش‌های مشاهده شده و گمشده Y هستند. برای سادگی فرض کنید که $r<n$ واحد نخست نمونه پاسخ داده‌اند و n واحد را با $i=1,\ldots,n$ شماره‌گذاری کرده‌ایم. به شرط نمونه‌گیری با احتمال مساوی می‌توان میانگین y را به‌عنوان میانگین پاسخ داده شده‌ها و واحدهای جانهی شده برآورد کرد. بنابراین، این جمله را به‌صورت زیر می‌نویسیم:

$$\bar{y}_{HD}=\{r\bar{y}_R+(n-r)\bar{y}_{NR}\}/n \qquad (۱۰)$$

که $\bar{y}_R$ میانگین واحدهای پاسخ داده شده است و

$$\bar{y}_{NR}=\sum_{i=1}^{r}\frac{H_i y_i}{n-r} \qquad (۱۱)$$

که H_i تعداد دفعاتی است که y_i برای یک مقدار گمشده Y جایگزین می‌شود و $\sum_{i=1}^{r}H_i=n-r$ تعداد واحدهای گمشده است. ویژگی‌های $\bar{y}_{HD}$ (میانگین نمونه با بهره‌مندی از روش بی‌درنگ) بستگی به روش بهره‌مندی شده برای تولید اعداد $\{H_1,\ldots,H_r\}$ دارد. ساده‌ترین نظریه وقتی به دست می‌آید که مقادیر جانهی شده را بتوان به‌عنوان انتخاب شده‌ها از مقادیر واحدهای پاسخ داده شده با بهره‌مندی از طرح نمونه‌گیری احتمالاتی در نظر گرفت، به طوری که توزیع $\{H_1,\ldots,H_r\}$ در کاربردهای تکرار شده روش بی‌درنگ شناخته شده باشد. میانگین و واریانس $\bar{y}_{HD}$ را می‌توان به‌صورت زیر نوشت:

$$\begin{aligned}&E(\bar{y}_{HD})=E[E(\bar{y}_{HD}\mid Y^{\text{obs}}]),\\&Var(\bar{y}_{HD})=Var[E(\bar{y}_{HD}\mid Y^{\text{obs}}]+E[Var(\bar{y}_{HD}\mid Y^{\text{obs}})]\end{aligned} \qquad (۱۲)$$

که امیدها و واریانس‌های داخل، روی $\{H_1,\ldots,H_r\}$ به شرط داده‌های مشاهده شده Y^{obs} هستند و امیدها و واریانس‌های بیرون، روی توزیع مدل Y یا توزیع نشانگرهای

نمونه‌گیری برای استنباط بر پایه طرح هستند. رابطه (۱۲)، واریانس اضافه شده از روش جانهی تصادفی را ذکر می‌کند. یکی از روش‌های جانهی بی‌درنگ عبارت است از جانهی بی‌درنگ با نمونه‌گیری تصادفی ساده بدون جایگذاری که در ادامه شرح داده می‌شود.

حال فرض کنید که جانهی با نمونه‌گیری تصادفی از شرکت کننده‌ها بدون جایگذاری باشد. برای تعریف روش، وقتی تعداد کمی شرکت کننده در جانهی نسبت به دریافت کننده‌ها وجود دارد، می‌نویسیم $n-r=kr+t$ که k یک عدد صحیح نامنفی و $0<t<r$ است. جانهی بی‌درنگ بدون جایگذاری، همه واحدهای ثبت شده را k بار انتخاب می‌کند و سپس t واحد اضافی به طور تصادفی بدون جایگذاری برای به دست آوردن $n-r$ مقدار لازم برای داده‌های گمشده انتخاب می‌کند. بنابراین، داریم:

$$\bar{y}_{NR}=(kr\bar{y}_R+t\bar{y}_t)/(n-r), \qquad (۱۳)$$

که $\bar{y}_t$ میانگین t مقدار اضافی Y است. اگر $\bar{y}_{HD2}$ نشان دهنده $\bar{Y}$ از این روش باشد، آنگاه:

$$E(\bar{y}_{HD2}\,|\,Y^{\text{obs}}]=\bar{y}_R,$$
$$Var(\bar{y}_{HD1}\,|\,Y^{\text{obs}})=(t/n)(1-t/r)s_{yR}^2/n \qquad (۱۴)$$

۴- مدل DEA ی BCC

n واحد تصمیم‌گیری را در نظر بگیرید که از نظر m ورودی و s خروجی بررسی می‌شوند. فرض کنید x_{ij} ($i=1,\ldots,m$) و y_{rj} ($r=1,\ldots,s$) مقادیر ورودی و خروجی آن‌ها برای $j=1,\ldots,n$ باشند. در مدل CCR فرض بر بازده به مقیاس ثابت است، به طوری که خروجی به همان نسبت ورودی افزایش می‌یابد. در حالت یک ورودی و یک خروجی، مرز تولید یک خط مستقیم است که از مبدأ می‌گذرد. در اقتصادهای تولیدی، به علت تأثیر ورودی‌های ثابت، بازده به مقیاس معمولاً در مرحله اولیه تولید، که در آن مقدار ورودی متغیر نسبتاً کوچک است، افزایشی است. به تدریج که مقدار ورودی متغیر افزایش می‌یابد، بازده به مقیاس با ثابت کاهش پیدا می‌کند و بالاخره کاهشی می‌شود. با در نظر گرفتن این پدیده، بنکر[۲۹] و همکاران [۲۲] مدل CCR را بسط دادند تا امکان بازده به مقیاس متغیر داشته باشد، که به آن مدل BCC گفته می‌شود. از نظر مفهومی، آن‌ها با در نظر گرفتن یک مقدار ثابت در تجمیع ورودی‌ها یا

خروجی‌ها، اجازه می‌دهند که مرز تولید از مبدأ فاصله بگیرد. مقدار ثابت در مرز تولید خطی نقش عرض از مبدأ را ایفا می‌کند. این مدل دو شکل ورودی و خروجی دارد.

۱-۴- مدل خروجی

بر خلاف مدل ورودی، که در آن مقدار کمینه ورودی‌های مورد نیاز برای تولید سطح معین خروجی برای اندازه‌گیری بیشینه مورد بهره‌مندی قرار می‌گیرد، مدل خروجی برای اندازه‌گیری کارایی، به دنبال مقدار بیشینه خروجی‌هایی است که می‌توان با مقدار داده شده ورودی‌ها تولید کرد. مدل ایجاد شده توسط بنکر و همکاران [۲۲] برای اندازه‌گیری کارایی از سمت خروجی عبارت است از:

$$\begin{aligned} \min \quad & \theta_o = \frac{\sum_{i=1}^{m} v_i x_{io} + u_o}{\sum_{r=1}^{s} u_r y_{ro}} \\ \text{s.t.} \quad & \theta_j = \frac{\sum_{i=1}^{m} v_i x_{ij} + u_o}{\sum_{r=1}^{s} u_r y_{rj} - u_o} \geq 1, \quad j = 1,\ldots,n, \\ & u_r, v_i \geq 0, \quad r = 1,\ldots,s; \quad i = 1,\ldots,m, \end{aligned} \qquad (۱۵)$$

u_o بدون محدودیت

تفاوت بین مدل (۱۵) و مدل CCR، یعنی مدل تحت شرایط بازده به مقیاس ثابت، در گنجاندن عرض از مبدأ u_o است. تابع هدف کسری خطی در مدل (۱۵) را می‌توان با تخصیص یک به مخرج و گذاشتن تابع هدف به‌عنوان صورت، خطی کرد. علت آن است که مدل (۱۵) جواب‌های متعدد دارد، بدان جهت که اگر $(\mathbf{u}^*, \mathbf{v}^*)$ یک جواب بهینه باشد، آنگاه $(c\mathbf{u}^*, c\mathbf{v}^*)$ نیز برای $c > 0$، جواب بهینه است. به این ترتیب، تخصیص دادن مقدار یک به مخرج برای کاهش درجه آزادی مقدار هدف بهینه، θ_o را تغییر نمی‌دهد، گرچه جواب بهینه $(\mathbf{u}^*, \mathbf{v}^*)$ ممکن است گوناگون باشد. قیود کسری خطی را به آسانی می‌توان با ضرب کردن هر دو طرف در مخرج خطی کرد، که منجر به مدل برنامه‌ریزی خطی زیر می‌شود:

$$\theta_o^* = \min \sum_{i=1}^{m} \upsilon_i x_{io} + \mu_o$$
$$\text{s.t.} \quad \sum_{i=1}^{m} \upsilon_i x_{ij} + \mu_o - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0, \quad j = 1, \ldots, n, \tag{۱۶}$$
$$\sum_{r=1}^{s} \mu_r y_{ro} = 1$$
$$u_r, v_i \geq 0, \quad r = 1, \ldots, s; \quad i = 1, \ldots, m,$$
μ_o بدون محدودیت

اگر مجموعه‌ای از وزن‌های مثبت μ_r^* $(r = 1, \ldots, s)$، υ_i^* $(i = 1, \ldots, m)$ و μ_o^* وجود داشته باشند تا $\theta_o^* = 1$ را تأمین کنند، آنگاه DMU_o کارای خوشبینانه یا به اختصار کارا نامیده می‌شود؛ در غیر این صورت به آن غیرکارای خوشبینانه می‌گویند.

۵- نمونه کاربردی

برای نشان دادن روش‌های پیشنهادی در مورد گنجاندن مقادیر گمشده در DEA، در این بخش یک ارزیابی از ۲۹ مدرسه متوسطه دولتی را واقع در آتنِ یونان ارائه می‌کنیم. این ۲۹ مدرسه که در کاربرد کنونی مورد بهره‌مندی قرار می‌گیرند، زیرمجموعه کوچکی از مدارس گنجانده شده را در بر می‌گیرند و به گونه‌ای انتخاب شده‌اند که مسئله مقادیر گمشده در DEA را بهتر نشان دهند. مجموعه داده‌ها از مقاله اسمیرلیس[۷] و همکاران [۲۳] گرفته شده است و در جدول ۱ نشان داده شده است.

مدل بهره‌مندی شده در این کاربرد کوشش می‌کند که کارایی مدارس، یعنی توانایی آن‌ها برای بهره‌مندی از منابع موجود (ورودی‌ها) جهت تولید حداکثر موفقیت آموزشی ممکن (خروجی) را اندازه‌گیری کند. کاربردهای مشابهی برای کارایی مدارس متوسطه دولتی برای کشورهای دیگر نیز انجام شده است [۲۴-۲۹]. در این کاربرد خاص، منابع مدرسه که در نظر گرفته شده‌اند، شامل بودجه/مخارج سالیانه، امکانات و تجهیزات مدرسه و وضعیت ذهنی و مهارت‌های دانش‌آموزان هستند. خروجی‌ها شامل تعداد دانش‌آموزان فارغ‌التحصیل شده که در دانشگاه‌ها و مؤسسات فنی پذیرفته شده‌اند، متوسطه نمره همه دانش‌آموزان در دروس

مدرسه و تعداد دانش‌آموزان فارغ‌التحصیل شده که عملکرد عالی داشته‌اند، هستند. ورودی‌ها و خروجی‌های مدل شامل موارد زیر هستند [۲۳]:

ورودی‌ها		
متغیر		شرح
x_1	بودجه	بودجه سالیانه مدرسه متشکل از میزان ارائه شده توسط وزارت آموزش و مقدار اهدا شده به‌عنوان کمک بلاعوض از طرف سازمان‌های محلی، اولیای دانش‌آموزان و غیره.
x_2	شاخص تأسیسات	شاخصی که نشان دهنده سطح تأسیسات و تجهیزات موجود در مدرسه است. این متغیر، وضعیت کلی ساختمان، تعداد کلاس‌ها، وجود سالن ورزشی، بدن‌سازی، آزمایشگاه‌های دروس فیزیک و شیمی، آزمایشگاه زبان، کتابخانه و نمونه‌های آن را خلاصه می‌کند. این شاخص در بخش خودآزمایی پرسشنامه گنجانده شده بود و مقدار آن مستقیماً توسط رئیس مدرسه تعیین می‌شد. دامنه این نمره بین ۱ (ساختمان بد، تعداد ناکافی کلاس‌های درس و امکانات و تجهیزات بسیار محدود) و ۱۰ (وضعیت عالی ساختمان و کلاس‌های درس، امکانات و تجهیزات کاملاً کافی).
x_3	سطح تحصیلات	درصد دانش‌آموزانی که والدین آن‌ها لااقل از یک مدرسه متوسطه فارغ‌التحصیل شده‌اند. این معیاری غیرمستقیمی از مهارت‌ها و وضعیت ذهنی و نگرش دانش‌آموزان در قبال آموزش است. پژوهش‌های انجام شده در خصوص آموزش داخل خانواده، نشان می‌دهد که دانش‌آموزانی که والدین آن‌ها لااقل در یک مدرسه متوسطه حضور یافته‌اند، اختلاف معنی‌داری از نظر موفقیت و عملکرد تحصیلی در مقایسه با سایر دانش‌آموزان دارند.
خروجی‌ها		
متغیر		شرح
y_1	پذیرش	تعداد دانش‌آموزان فارغ‌التحصیل مدرسه که موفق به پذیرش در امتحانات ملی ورود به دانشگاه و مؤسسات فنی یونان شده‌اند.
y_2	معدل نمره	معدل نمرات سالیانه دانش‌آموزان فارغ‌التحصیل در دروس مدرسه. مقدار آن بین ۱۰ تا ۲۰ است.
y_3	دانش‌آموزان ممتاز	تعداد دانش‌آموزان فارغ‌التحصیل که به نمره متوسط عالی دست یافته‌اند (بین ۱۸ و ۲۰).

در کاربرد فعلی، با توجه به اینکه عملکرد دانش‌آموزان مستقیماً متناسب با ورودی‌ها (بودجه، امکانات و مهارت‌های دانش‌آموزان) نیست، لذا بازده به مقیاس متغیر را در نظر می‌گیریم و از مدل BCC (۱۶) بهره‌مندی می‌کنیم. کارهای پژوهشی پیشین در زمینه کارایی مدارس نیز نشان دهنده همین مطلب است [۲۳، ۲۴]. به علاوه، مدل با ماهیت خروجی انتخاب می‌شود، زیرا تنها خروجی‌ها به وسیله مدارس قابل کنترل هستند.

در مجموعه مدارسی که در ارزیابی گنجانده می‌شوند، تعدادی از مدارس به علت کامل نبودن پرونده دانش‌آموزان و فقدان امکانات دفترداری مناسب، قادر نبودند که داده‌های لازم برای برخی از ورودی‌ها و خروجی‌ها را ارائه کنند. بالاخص، مدارس ۷ و ۱۹ مقادیر گمشده در متغیر ورودی «بودجه» داشتند، مدارس ۴ و ۲۹ داده‌های مربوط به متغیر ورودی «سطح تحصیلات» را ارائه نکردند و مدارس ۶، ۷، ۱۰ و ۱۹ برای متغیر خروجی «معدل نمره» مقادیر گمشده داشتند. این موارد از مقادیر گمشده با روش‌های گوناگونی برآورد شدند، که در ادامه بیان شده است.

جدول ۱: مجموعه داده‌های ۲۹ مدرسه متوسطه دولتی.

مدرسه	ورودی‌ها			خروجی‌ها		
	بودجه x_1	شاخص تأسیسات x_2	سطح تحصیلات x_3	پذیرش y_1	معدل نمره y_2	دانش‌آموزان ممتاز y_3
۱	۲۳۹۴۰	۶	۵۲/۱۷	۱۹	۱۴/۷	۱۰
۲	۲۵۴۵۰	۵	۷۶/۴۳	۳۸	۱۴/۷	۱۴
۳	۲۴۰۰۰	۴	۴۳/۰۰	۳۴	۱۵/۰	۴
۴	۲۶۵۰۰	۷		۲۹	۱۴/۳	۴
۵	۳۱۲۰۰	۶	۴۳/۷۰	۴۸	۱۴/۰	۱۱
۶	۳۲۶۰۰	۵	۷۶/۴۳	۳۶		۱۷
۷		۵	۵۲/۲۱	۷۳		۱۸
۸	۳۵۶۰۰	۵	۹۳/۶۷	۴۰	۱۵/۷	۲۲
۹	۳۹۱۶۰	۴	۹۶/۱۷	۳۳	۱۵/۱	۳۸
۱۰	۴۲۸۰۰	۴	۴۳/۸۰	۶۲		۱۳
۱۱	۴۲۸۴۰	۷	۸۲/۴۳	۷۸	۱۴/۵	۲۷
۱۲	۴۱۰۰۰	۴	۷۵/۱۷	۶۲	۱۳/۶	۲۷
۱۳	۴۵۹۸۰	۷	۸۱/۹۶	۷۰	۱۵/۲	۲۸
۱۴	۵۱۰۰۰	۷	۷۶/۴۳	۵۹	۱۵/۵	۱۵
۱۵	۵۲۲۰۰	۵	۴۳/۲۰	۷۶	۱۵/۷	۲۵
۱۶	۵۶۰۰۰	۷	۵۴/۷۱	۵۶	۱۳/۲	۲۶
۱۷	۵۶۷۰۰	۷	۷۵/۱۷	۵۹	۱۳/۳	۳۳
۱۸	۵۸۱۴۰	۴	۳۷/۷۹	۷۸	۱۴/۸	۳۴
۱۹		۴	۵۹/۴۰	۹۶		۱۸
۲۰	۶۰۱۰۰	۷	۷۸/۶۷	۹۵	۱۵/۵	۳۵

مدرسه	ورودی‌ها			خروجی‌ها		
	بودجه x_1	شاخص تأسیسات x_2	سطح تحصیلات x_3	پذیرش y_1	معدل نمره y_2	دانش‌آموزان ممتاز y_3
۲۱	۶۰۰۴۰	۷	۴۷/۵۶	۸۳	۱۴/۵	۲۳
۲۲	۶۳۴۵۰	۷	۵۸/۸۶	۷۶	۱۴/۲	۴۹
۲۳	۶۱۱۱۰	۷	۵۶/۲۴	۹۸	۱۱/۰	۳۳
۲۴	۶۱۸۲۰	۷	۶۸/۱۲	۸۵	۱۴/۴	۳۳
۲۵	۶۵۰۰۰	۵	۵۸/۸۶	۴۷	۱۳/۴	۳۶
۲۶	۶۴۰۵۰	۷	۷۶/۲۲	۶۸	۱۴/۹	۴۶
۲۷	۷۴۶۰۰	۶	۳۹/۰۰	۱۱۱	۱۵/۳	۳۹
۲۸	۷۶۶۵۰	۵	۵۶/۲۴	۱۲۴	۱۳/۲	۴۸
۲۹	۸۲۷۴۰	۴		۱۰۰	۱۳/۷	۳۷

- برآورد مقادیر گمشده با بهره‌مندی از روش آنتروپی

جدول ۲ برآورد مقادیر گمشده با بهره‌مندی از روش آنتروپی را نشان می‌دهد.

جدول ۲: برآورد مجموعه داده‌های گمشده با بهره‌مندی از روش آنتروپی.

مدرسه	ورودی‌ها		خروجی
	بودجه x_1	سطح تحصیلات x_3	معدل نمره y_2
۴		۶۶/۴۷	
۶			۱۴/۴
۷	۵۳۰۵۶/۸		۱۴/۴
۱۰			۱۴/۴
۱۹	۵۳۰۵۶/۸		۱۴/۴
۲۹		۶۶/۴۷	

- برآورد مقادیر گمشده با بهره‌مندی از روش EM

جدول ۳ برآورد مقادیر گمشده با بهره‌مندی از روش EM را نشان می‌دهد.

جدول ۳: برآورد مجموعه داده‌های گمشده با بهره‌مندی از روش EM.

مدرسه	ورودی‌ها		خروجی
	بودجه x_1	سطح تحصیلات x_3	معدل نمره y_2
۴		۶۶/۲۸	
۶			۱۴/۷۸
۷	۵۱۸۸۰/۵		۱۴/۲۴
۱۰			۱۴/۳۳
۱۹	۵۰۸۸۹/۴۷		۱۴/۳۲
۲۹		۵۸/۹	

- برآورد مقادیر گمشده با بهره‌مندی از روش رگرسیون

جدول ۴ برآورد مقادیر گمشده با بهره‌مندی از روش رگرسیون را نشان می‌دهد.

جدول ۴: برآورد مجموعه داده‌های گمشده با بهره‌مندی از روش رگرسیون.

مدرسه	ورودی‌ها		خروجی
	بودجه x_1	سطح تحصیلات x_3	معدل نمره y_2
۴		۸۱/۹۶	
۶			۱۴/۸
۷	۳۹۱۶۰		۱۴
۱۰			۱۵/۵۰
۱۹	۴۵۹۸۰		۱۴/۷۰
۲۹		۴۳	

- برآورد مقادیر گمشده با بهره‌مندی از روش جانهی با میانگین

جدول ۵ برآورد مقادیر گمشده با بهره‌مندی از روش جانهی با میانگین را نشان می‌دهد.

جدول ۵: برآورد مجموعه داده‌های گمشده با بهره‌مندی از روش جانهی با میانگین.

مدرسه	ورودی‌ها		خروجی
	بودجه x_1	سطح تحصیلات x_3	معدل نمره y_2
۴		۶۴/۱	
۶			۱۴/۳۶

۷	۵۰۱۷۳		۱۴/۳۶
۱۰			۱۴/۳۶
۱۹	۵۰۱۷۳		۱۴/۳۶
۲۹		۶۴/۱	

با جانهی مقادیر برآورد شده به جای مقادیر گمشده در جدول ۱، کارایی ۲۹ مدرسه را با اجرای مدل BCC (۱۶) به دست می‌آوریم که در جدول ۶ نشان داده شده‌اند. در هر چهار روش جانهی، مدارس ۱، ۲، ۳، ۵، ۸، ۹، ۱۱، ۱۲، ۱۵، ۱۸، ۱۹، ۲۰، ۲۲، ۲۶، ۲۷، ۲۸ و ۲۹ به‌صورت کارا شناسایی شده‌اند. همچنین، در جانهی به روش رگرسیون، مدارس ۷ و ۱۰ نیز کارا شناسایی شده‌اند. به این ترتیب، مدارس ۱۹ و ۲۹ همواره کارا هستند، در حالی که مدرسه ۴ همیشه غیرکارا است.

جدول ۶: کارایی‌های به دست آمده از روش‌های گوناگون روش جانهی.

مدرسه	کارایی			
	با جانهی به روش آنتروپی	با جانهی به روش EM	با جانهی به روش رگرسیون	با جانهی به روش میانگین
۱	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۲	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۳	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۴	۱/۰۵۹۵	۱/۰۵۹۵	۱/۰۵۹۵	۱/۰۵۹۵
۵	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۶	۱/۰۶۸	۱/۰۴۳۴	۱/۰۴۲۱	۱/۰۷۰۴
۷	۱/۰۸۲۱	۱/۰۸۶۱	۱/۰۰۰۰	۱/۰۶۹۹
۸	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۹	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۱۰	۱/۰۰۸۹	۱/۰۱۸۱	۱/۰۰۰۰	۱/۰۲۱۲
۱۱	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۱۲	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۱۳	۱/۰۰۰۷	۱/۰۰۰۷	۱/۰۰۰۷	۱/۰۰۰۷
۱۴	۱/۰۱۲۹	۱/۰۱۲۹	۱/۰۱۲۹	۱/۰۱۲۹
۱۵	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۱۶	۱/۱۷۶۶	۱/۱۷۶۶	۱/۱۷۶۶	۱/۱۷۶۶

۱۷	۱/۱۴۷۹	۱/۱۴۷۹	۱/۱۴۷۹	۱/۱۴۷۹
۱۸	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۱۹	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۲۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۲۱	۱/۰۷۲۶	۱/۰۷۲۶	۱/۰۷۲۶	۱/۰۷۲۶
۲۲	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۲۳	۱/۰۲۸۴	۱/۰۳۵۳	۱/۰۶۵۲	۱/۰۳۷۷
۲۴	۱/۰۷۲۷	۱/۰۷۲۷	۱/۰۷۲۷	۱/۰۷۲۷
۲۵	۱/۱۰۴۷	۱/۱۰۴۷	۱/۱۰۴۷	۱/۱۰۴۷
۲۶	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۲۷	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۲۸	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰
۲۹	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰

۶- نتیجه‌گیری

در بسیاری از کاربردها، ماتریس‌های ورودی و خروجی شامل درایه‌های خالی هستند که قبل از تحلیل DEA کنارگذاری می‌شوند. این پژوهش نخستین کوشش سیستماتیک با بهره‌مندی از رویکردهای آماری برای اعمال DEA بر داده‌های نامتوازن با مقادیر گمشده بود. یافته‌های ما دلگرم کننده است، بدین معنا که ظاهراً DEA ی استاندارد با تغییراتی اندک و ساده، قادر به کار کردن با درایه‌های خالی هست. در این مقاله توجه کردیم که با اختصاص مقادیر مناسب به درایه‌های گمشده، می‌توانیم مدل عادی DEA را به طور خودکار اجرا کنیم. فکر می‌کنیم که این روش‌ها می‌توانند توانایی ما را برای بهره‌مندی از DEA در مواردی که داده‌ها با مشکل پوشش مواجه هستند، افزایش دهند.

علیرغم این نتایج دلگرم کننده، این بررسی صرفاً یک قدم ابتدایی در یک پژوهش ادامه‌دار است. با آنکه رویکرد معرفی شده در این پژوهش می‌تواند قابلیت‌های ما را برای به کار بردن DEA در موقعیت‌هایی که پوشش داده‌ها مشکل ایجاد می‌کند، افزایش دهد ولی راه‌حل‌های فنی این پژوهش را نباید درمانی برای همه مشکلات داده‌ها به شمار آورد. مشکل اصلی آن است که منصفانه نیست که DMU هایی را که درایه‌های گمشده زیادی دارند، با DMU هایی که داده‌های کامل دارند، مقایسه کنیم.

با این حال، در بسیاری از موارد، منصفانه به نظر می‌رسد که DMU هایی را که داده‌های

خود را به طور عمومی گزارش می‌کنند، پاداش بدهیم و آن‌ها را بر پایه تعداد بیشتری ورودی و خروجی ارزیابی کنیم. این امر می‌تواند DMU ها را تشویق کند که در آینده به گزارش‌دهی داده‌ها توجه بیشتری کنند که این امر خصوصاً در مقایسه‌های بین‌المللی اهمیت زیادی دارد. ابعاد راهبردی نحوه مواجهه با داده‌های گمشده در سنجش کارایی، مشوق‌های صحیحی برای گزارش‌دهی داده‌های واقعی ارائه می‌کند که شایسته پژوهش‌های بیشتر است.

یک مسیر جالب دیگر برای پژوهش‌ها این است که کوشش کنیم تأثیر ابعاد را بر نمره کارایی بررسی کنیم. این نکته شناخته شده است که اگر یک ورودی-خروجی دیگر اضافه شود، نمره کارایی نمی‌تواند کاهش یابد ولی درباره میزان افزایش نمی‌توانیم چیز زیادی بگوییم. اگر ضرر ناشی از ابعاد کوچک‌تر را بتوان به‌صورتی معنی‌دار و یکسان اندازه‌گیری کرد، شاید بتوانیم نمرات کارایی را از نظر اختلاف ابعاد اصلاح کنیم و لذا مسئله مقایسه غیرمنصفانه را کمتر کنیم یا آنکه کاملاً از بین ببریم.

امید می‌رود که رویکرد پیشنهادی بتواند بینش‌های بیشتری برای اندازه‌گیری کارایی ارائه کند و نظریه و رویکرد DEA را تقویت کند. از کارهای تحقیقاتی آینده این است که ببینیم آیا رویکرد را می‌توان بسط داد تا DEA ی نادقیق [۳۰، ۳۱] را نیز در بر بگیرد. البته هنوز چیزهای خوب زیاد دیگری مانند ارزیابی همزمان از دو دیدگاه خوشبینانه و بدبینانه [۳۲-۳۵] وجود دارد، که می‌توان آن‌ها را بررسی کرد. خوانندگان علاقه‌مند می‌توانند خودشان این مسائل را بررسی کنند.

۷- پی‌نوشت‌ها

۱. Data envelopment analysis (DEA).
۲. Post.
۳. Kuosmanen.
۴. Simar.
۵. Wilson.
۶. Griliches.
۷. Kao.
۸. Liu.
۹. Gardijan.
۱۰. Lukač.
۱۱. Chen.
۱۲. Duarte.
۱۳. Stead.
۱۴. Wheat.
۱۵. Little.
۱۶. Rubin.
۱۷. Missing Completely at Random (MCAR).
۱۸. Missing at Random (MAR).
۱۹. Missing not at random (MNAR).
۲۰. Fleiss.
۲۱. Song.
۲۲. Shepperd.
۲۳. Cohen.
۲۴. Sufficient statistics.

۲۵. Catellier.
۲۶. Tanner.
۲۷. Wong.
۲۸. Gibbs.
۲۹. Banker.
۳۰. Smirlis.

۸- منابع

[1] Charnes, A., Cooper, W.W., Rhodes, E. Measuring the efficiency of decision making units, *European Journal of Operational Research*, 2, 1978, 429–444.

[2] Post, T., Cherchye, L., Kuosmanen, T. Nonparametric efficiency estimation in stochastic environments, *Operations Research*, 50(4), 2002, 645–655.

[3] Kuosmanen, T., Post, T., Scholtes, S. Non-parametric tests of productive efficiency with errors-in-variables. *Journal of Econometrics*, 136(1), 2007, ۱۳۱–۱۶۲.

[4] Simar, L., Wilson, P. Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis*, 13(1), 2000, 49–78.

[5] Griliches, Z. *Economic data issues*. In: Griliches Z and Intriligator MD (eds). *Handbook of Econometrics*, Vol. III, Chapter 25. Elsevier: Amsterdam/New York, 1986.

[6] Kao, C., Liu, S.-T. Data envelopment analysis with missing data: An application to University libraries in Taiwan, *Journal of the Operational Research Society*, ۵۱(۸), ۲۰۰۰, ۸۹۷–۹۰۵.

[7] Gardijan, M., Lukač, Z. Measuring the relative efficiency of the food and drink industry in the chosen EU countries using the data envelopment analysis with missing data, *Central European Journal of Operations Research*, 26, 2018, ۶۹۵–۷۱۳.

[8] Chen, C., Ren, J., Tang, L., Liu, H. Additive integer-valued data envelopment analysis with missing data: A multi-criteria evaluation approach, *PloS one*, ۱۵(۶), ۲۰۲۰, ۰۲۳۴۲۴۷.

[9] Duarte, L.T., Mussio, A.P., Torezzan, C. Dealing with missing information in data envelopment analysis by means of low-rank matrix completion, *Annals of Operations Research*, 286, 2020, 719–732.

[10] Stead, A.D., Wheat, P. The case for the use of multiple imputation missing data methods in stochastic frontier analysis with illustration using English local highway data, *European Journal of Operational Research*, 280(1), 2020, 59-

۷۷.

[11] Little, R.J.A., Rubin, D.B. Statistical Analysis with Missing Data. New York: Wiley, 1987.

[12] Fleiss J.L., Levin B., Paik M.C. Statistical Methods for Rates and Proportions. ۲□□ □□. □□□ □□□□: □□□□ □□□□□ □ □□□□, ۲۰۰۲.

[13] Ibrahim, J.G., Chen, M.H., Lipsitz, S.R. Bayesian methods for generalized linear models with covariates missing at random, *Canadian Journal of Statistics*, 30(1), 2002, 55–78.

[14] Karimlou, M., Jandaghi, G.R., Mohammad, K., Wolfe, R., Azam, K. A comparison of parameter estimates in standard logistic regression using WinBUGS MCMC and MLE methods in R for different sample sizes, *Far East Journal of Theoretical Statistics*, 19(2), 2006, 281–292.

[15] Rubin, D.B. Multiple Imputation after 18+ Years, *Journal of the American Statistical Association*, 91, 1996, 473–489.

[16] Little, R.J A., Rubin, D.B. Statistical Analysis with Missing Data, John Wiley and Sons, 2002.

[17] Cohen, M.P. A new approach to imputation, *American Statistical Association Proceedings of the Section on Survey Research Methods*, 1996, 293–298.

[18] Song, Q., Shepperd, M. Missing data imputation techniques, *International Journal of Business Intelligence and Data Mining*, 2(3), 2007, 261–291.

[19] Dempster, A.P., Laird, N.M., Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B*, 1977, 1–38.

[20] Catellier, D.J., Hannan, P.J., Murray, D.M., Addy, C.L., Conway, T.L., Yang, S., Rice, J.C. Imputation of missing data when measuring physical activity by accelerometry, *Medicine and science in sports and exercise*, 37 (11 Suppl), ۲۰۰۵, ۵۵۵–۵۶۲.

[21] Tanner, M.A., Wong W.H. The calculation of posterior distribution by dats augmentation (with discussion), *Journal of the American Statistical Association*, 82(398), 1987, 528–550.

[22] Banker, R.D., Charnes, A., Cooper, W.W. Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science*, 30, 1984, 1078–1092.

[23] Smirlis, Y.G. Maragos, E.K. and Despotis, D.K. Data envelopment analysis

with missing values: An interval DEA approach, *Applied Mathematics and Computation*, 2006, 177, 1–10.

[24] Conceicao Silva Portela, M.A., Thanassoulis, E. Decomposing school and school-type efficiency, *European Journal of Operational Research*, 132, 2001, ۳۵۷–۳۷۳.

[25] Bradley, S., Johnes, G., Millington, J. The effect of competition of secondary schools in England, *European Journal of Operational Research*, 135, 2001, ۵۴۵–۵۶۸.

[26] Kirjavainen, T., Loikkanen, H. Efficiency differences of Finnish Senior secondary schools: An application of DEA and Tobit analysis, *Economics of Education Review*, 1998, 17, 377–394.

[27] Soteriou, A., Karahana, E., Papanastasiou, C., Diakourakis, M. Using DEA to evaluate the efficiency of secondary schools: The case of Cyprus, *International Journal of Educational Management*, 12, 1998, 65–73.

[28] Maragos, E.K., Despotis, D.K. The evaluation of the efficiency with data envelopment analysis in case of missing values: A fuzzy approach, *WSEAS Transactions on Mathematics*, 3(3), 2004, 656–663.

[29] Muñiz, M.A. Separating managerial inefficiency and external conditions in data envelopment analysis, *European Journal of Operational Research*, 143(3), ۲۰۰۲, ۶۲۵–۶۴۳.

[30] Azizi, H., Amirteimoori, A., Kordrostami, S. Measurement of the worst practice of decision-making units: Incorporating both undesirable outputs and non-discretionary inputs into imprecise DEA, *Modern Researches in Decision Making*, 3(2), 2018, 197-222. (In Persian)

[31] Azizi, H., Amirteimoori, A., Kordrostami, S. A data envelopment analysis approach with efficient and inefficient frontiers for supplier selection in the presence of both undesirable outputs and imprecise data, *Modern Researches in Decision Making*, 1(2), 2016, 139-170. (In Persian)

[32] Azizi, H. Efficiency assessment in data envelopment analysis using efficient and inefficient frontiers, *Management Research in Iran*, 16(3), 2012, 153–۱۷۳. (In Persian)

[33] Azizi, H., Jahed, R. Supplier Selection in Volume Discount Environments in the Presence of Both Cardinal and Ordinal Data: A New Approach Based On Double Frontiers DEA, *Management Research in Iran*, 19(3), 2015, 191–217.

(In Persian)

[34] Azizi, H., Amirteimoori, A. Flexible Measures in Production Process: A New Approach Based On Double-Frontier DEA, *Modern Researches in Decision Making*, 2(2), 2017, 197-216. (In Persian)

[35] Azizi, H. New models for selecting third-party reverse logistics providers in the presence of multiple dual-role factors: Data envelopment analysis with double frontiers, *Decisions and Operations Research*, 5(2), 2020, 221-232. (In Persian)