

مروری بر داده‌های گم‌شده

الهه کاظمی^۱، *مسعود کریملو^۲، مهدی رهگذر^۲

Missing Data

Kazemi E. (M.Sc.)¹, *Karimlo M. (Ph.D.)², Rahgozar M. (Ph.D.)²

Abstract

In this paper, we are presenting the basic concepts of missing data in a very simple but practical approach.

Missing data are ubiquitous throughout the social, behavioral, and medical sciences. In Statistics, missing data occur when no data value is stored for the variable in the current observation. Missing data reduce the representativeness of the sample and can therefore distort inferences about the population.

Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. All the methods of parameters estimation are based on the completion of data set assumption and only in this case the result will be a non-biased one, and with the increase of missing proportion, the rate of biased results increase too.

For decades, researchers have relied on a variety of old techniques that attempt to "fix" the data by discarding incomplete cases or by filling in the missing values. Unfortunately, most of these techniques require a relatively strict assumption about the cause of missing data and are prone to substantial bias.

Keywords: Missing Data, Missing Completely at Random, Missing at Random, Non Ignorable, Missing by Natural Design

چکیده

در این مقاله سعی شده است که مفاهیم گم‌شدگی داده به صورت ساده و کاربردی توضیح داده شود.

گم‌شدگی داده در تمامی پژوهش‌های علوم اجتماعی، رفتاری، پزشکی وجود دارد. در آمار، گم‌شدن داده به وضعیتی گفته می‌شود که بخشی از مجموعه داده‌ها گزارش نشده باشند. گم‌شدگی داده باعث کاهش تطابق جامعه نمونه با جامعه کل شده و می‌تواند منجر به نتیجه‌گیری اشتباه در مورد جمعیت اصلی شود.

گم‌شدگی داده یک اتفاق معمول بوده و بسته به میزان آن، می‌تواند اثر قابل توجهی در نتیجه‌گیری به دست آمده از داده‌ها داشته باشد. تمامی روش‌های برآورد پارامترها بر پایه فرض کامل بودن مجموعه داده‌ها استوار است و تحت برقراری این شرایط منجر به برآوردهایی نارایب می‌شوند؛ و البته با افزایش نسبت گم‌شدگی، مقدار اریبی نیز افزایش خواهد یافت.

برای دهه‌ها، محققین از روش‌های قدیمی استفاده می‌کرده‌اند، این روش‌ها متکی به تصحیح مجموعه داده‌ها با صرف‌نظر کردن از موردهای دارای مقادیر گم‌شده و یا جایگزینی مقادیری تخمینی با مقادیر گم‌شده بودند. متأسفانه اکثر این روش‌ها وابسته به برقراربودن فرض دلایل گم‌شدگی داده و نوع سازوکار گم‌شدگی است؛ و در صورت عدم برقراری این فرض منجر به اریبی نتایج می‌شود.

کلیدواژه‌ها: داده گم‌شده، سازوکار گم‌شدگی کاملاً تصادفی، سازوکار گم‌شدگی تصادفی، سازوکار گم‌شدگی غیر قابل اغماض، سازوکار گم‌شدگی به علت ذات طرح

۱. کارشناس ارشد آمار حیاتی، دانشگاه علوم بهزیستی و توانبخشی؛ ۲. دکترای آمار حیاتی، دانشیار دانشگاه علوم بهزیستی و توانبخشی

*آدرس نویسنده مسئول: تهران، اوین، بلوار دانشجو، خیابان کودکیار، دانشگاه علوم بهزیستی و توانبخشی، گروه آموزشی آمار حیاتی؛ تلفن: ۲۲۱۸۰۱۴۶؛ رایانامه:

mkarimlo@yahoo.com

1. M.Sc. of Biostatistics, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran; 2. Biostatistician, Associate Professor of University of Social Welfare and Rehabilitation Sciences, Tehran, Iran

*Correspondent Address: Statistics and Computer Department, University of Social Welfare and Rehabilitation Sciences, Koodakyar Alley, Daneshjoo Blv., Evin, Tehran, Iran; Tel: +98 21 22180 146; E-mail: mkarimlo@yahoo.com

مقدمه

علی رغم این که در اکثر مثال‌های کتاب‌های درسی داده‌ها به طور کامل موجود می‌باشند، جمع آوری داده‌ها به طور کامل در تحقیقات عملی امکان پذیر نمی‌باشد. داده گم شده یک مشکل عمومی در تحقیقات علوم پزشکی، اپیدمیولوژیکی، توانبخشی، اجتماعی و رفتاری می‌باشد و عملاً تجزیه و تحلیل آماری را به سوی نتایج اریب سوق داده و نهایتاً دستیابی به یک نتیجه گیری مفید از داده‌های جمع آوری شده را با مشکل مواجه می‌سازد. با وجود تمامی این مشکلات گم‌شدگی بهتر از جواب اشتباه می‌باشد.

در ادبیات آماری اصطلاحات مختلف و غالباً مترادفی برای این مفهوم وجود دارد. این اصطلاحات عبارتند از مقادیر گم شده^۱، داده‌های گم شده^۲، داده‌های ناقص^۳ و بی پاسخ^۴. گم‌شدگی هم در متغیر پاسخ و هم در متغیرهای مستقل رخ می‌دهد. دو نوع گم‌شدگی (بی پاسخی) داریم، بی پاسخی آزمودنی^۵ که تمامی اطلاعات برای یک نمونه گم شده باشد (آزمودنی گم شده است). بی پاسخی در یک سؤال^۶ که برخی از اطلاعات یک نمونه موجود نباشد (گم شده باشد).

به علاوه گم‌شدن در متغیرها به صورت‌های مختلفی می‌تواند رخ دهد. به عنوان مثال برخی از افراد شرکت‌کننده در مطالعه از ادامه همکاری انصراف می‌دهند، یا از پاسخ دادن به برخی از سؤالات اجتناب می‌کنند. محققین، تکنسین‌ها، جمع آوری کننده داده‌ها ممکن است اشتباهاتی را انجام دهند. در برخی مطالعات گم‌شدگی ممکن است به علت نوع طرح پژوهشی رخ داده باشد، به طور مثال در مطالعه‌ای برای جمع آوری داده‌های مورد نظر یک نمونه‌گیری دو مرحله‌ای انجام می‌شود، در مرحله اول مقادیر متغیرهایی که اندازه‌گیری آنها آسان و ارزان است برای تمامی افراد شرکت‌کننده در مطالعه جمع آوری می‌شود و سپس در مرحله دوم مقادیر متغیرهایی که اندازه‌گیری آنها پرهزینه و پیچیده می‌باشد برای زیر مجموعه‌ای از افراد شرکت‌کننده در مطالعه جمع آوری می‌شود، پس بنابراین جزئیات برای کل افراد شرکت‌کننده در مطالعه موجود نمی‌باشد. در یک مطالعه گذشته نگر به علت نقص مدارک و سوابق ممکن

است برخی از اطلاعات در دسترس نباشد. همچنین این احتمال هست که به علت نقص یا ضعف دستگاه و تجهیزات، امکان مشاهده و اندازه‌گیری وجود نداشته باشد و یا در بعضی از مطالعات نظر سنجی، افرادی قادر به اظهار نظر دقیق نباشند؛ و یا به هر دلیلی داده جمع آوری شده برای برخی از افراد شرکت‌کننده در مطالعه گم شود.

آن چه که در قدم اول در برخورد با داده‌های گم شده حائز اهمیت است بازنگری و مشاهده مجدد آزمودنی‌های مورد مطالعه و اقدام به تکمیل مقادیر گم شده است (۳-۱). به طور کلی سه روش در نحوه بررسی داده‌های گم شده مورد استفاده قرار می‌گیرد (۴):

۱. روش‌های مبتنی بر واحدهای کامل^۷

۲. روش‌های مبتنی بر جانپ^۸

۳. روش‌های مبتنی بر مدل

هر روشی را که برای تحلیل داده‌های گم شده به کار می‌بریم نقاط ضعف و قوت خودش را دارد، که وابسته است به:

۱. نسبت و الگوی گم‌شدگی

۲. مدلی که برای تحلیل پرسش‌های تحقیق انتخاب شده است

۳. تعداد متغیرهای مورد استفاده در تحقیق

۴. نوع متغیرهایی که دارای مقادیر گم شده می‌باشند

۵. سازوکار^۹ گم‌شدگی

امروزه بسیاری از محققین روش‌هایی را برای مقابله با مشکل داده گم شده ارائه کرده‌اند که اکثر این روش‌ها وابسته به نوع سازوکار گم‌شدگی داده‌ها می‌باشند؛ لذا باید نوع سازوکار گم‌شدن معلوم باشد تا بتوان روش تحلیل مناسبی را انتخاب نمود. متأسفانه برخی از محققین بین الگوی گم‌شدگی و سازوکار گم‌شدگی تمایزی قائل نمی‌باشند و این دو تعریف را به جای هم استفاده می‌کنند که اشتباه است. یک الگوی گم‌شدگی داده نشان دهنده ترکیب داده‌های گم شده و مشاهده شده در مجموعه داده‌ها می‌باشد، در حالی که سازوکار گم‌شدگی داده‌ها، رابطه ممکن بین مقادیر گم شده و مشاهده شده را توضیح می‌دهد (۳). به همین دلیل در ادامه در رابطه با الگوی گم‌شدگی و سازوکار گم‌شدگی توضیحات بیشتری را ارائه می‌دهیم.

- | | |
|----------------------|----------------------|
| 1. Missing Values | 2. Missing Data |
| 3. Incomplete Data | 4. Non Response |
| 5. Unit Non Response | 6. Item Non Response |

7. Complete Cases ,Complete Records
8. Imputation 9. Mechanism

مقایسه‌های آزمون تی: در این روش مواردی را که مقادیر گم شده و مشاهده شده دارند را برای یک متغیر خاص جدا می‌کنیم و از آزمون تی برای مقایسه میانگین‌های این دو گروه بر روی سایر متغیرها استفاده می‌کنیم. سازوکار گم‌شدگی کاملاً تصادفی بیان می‌کند که میانگین موارد مشاهده شده و گم شده باید برابر باشند. قبول فرض صفر این آزمون منجر به پذیرش سازوکار کاملاً تصادفی داده‌ها می‌شود، و رد فرض صفر، معنی دار شدن آماره تی یعنی تفاوت زیاد میانگین‌ها، بیان می‌کند که مجموعه داده‌ها ممکن است دارای سازوکار گم‌شدگی تصادفی یا غیر قابل اغماض باشند (۷).

آزمون گم‌شدگی کاملاً تصادفی لیتل: لیتل (۹) یک آزمون تی که برای چند متغیره بسط داده شده است را پیشنهاد کرده که به طور همزمان تفاوت میانگین‌ها را برای تمامی متغیرهای مجموعه داده‌ها ارزیابی می‌کند. مانند آزمون تی یک متغیره، آزمون لیتل نیز تفاوت میانگین‌ها را در بین زیر گروه‌ها مقایسه می‌کند. فرض صفر در این آزمون وجود سازوکار گم‌شدگی کاملاً تصادفی داده‌ها می‌باشد. این آزمون در نرم افزارهای آماری مانند SPSS موجود می‌باشد (۱۳).

۲. گم شدن تصادفی^۱ (MAR): در این حالت گم‌شدگی در متغیر، X_{mis} بستگی به متغیری که برای همه افراد شرکت‌کننده در مطالعه به طور کامل مشاهده شده است، X_{obs} دارد ولی به خود متغیر، X_{mis} بستگی ندارد. این سازوکار آزمون پذیر نمی‌باشد (۱۴، ۳).

برای تشخیص سازوکار گم‌شدگی تصادفی فلایس و همکاران^۲ در کتاب خود (۲) روشی را ارائه کرده‌اند. این روش عمومی نبوده و زمانی کاربرد دارد که مجموعه داده‌ها به گونه‌ای باشد که بتوان از رگرسیون لجستیک استفاده کرد و همچنین متغیری که دارای مقادیر گم شده، X_{mis} ، می‌باشد متغیری رسته‌ای و دو حالتی با شد. متغیرهای X_{obs} و X_{mis} متغیر مستقل می‌باشند و متغیر Y را به عنوان متغیر وابسته در نظر می‌گیریم. شیوه کار به صورت الگوریتم زیر می‌باشد:

۱. متغیر نشانگر گم‌شدگی دو حالتی Δ_i را به صورت زیر تعریف می‌کنیم:

$$\Delta_i = \begin{cases} 1 & \text{اگر } X_{mis(i)} \text{ مشاهده شده باشد} \\ \text{صفر} & \text{اگر } X_{mis(i)} \text{ گم شده باشد} \end{cases}$$

۲. مجموعه داده‌های بسط داده شده را به وسیله تکثیر کردن هر مقدار گم شده با جایگذاری $X_{mis(i)}=0$ و $X_{mis(i)}=1$ درست می‌کنیم. مجموعه داده بسط داده شده جدید شامل n' واحد می‌باشد. برای داده‌های جدید سایر متغیرها، Y_i و $X_{obs(i)}$ ، را به همان صورت که مشاهده شده‌اند تکرار می‌کنیم. در این مجموعه داده مقدار n' از فرمول زیر به دست می‌آید:

$$n' = \sum_{i=1}^n \Delta_i + 2 \sum_{i=1}^n (1 - \Delta_i)$$

۳. مقادیر اولیه‌ای را برای α ها در نظر می‌گیریم، با استفاده از این مقادیر اولیه برای هر واحد وزن‌های، w_i ، را برای $i=1,2,\dots,n'$ باید محاسبه کنیم. برای داده‌های مشاهده شده، $w_i=1$ می‌باشد. برای مقادیر تکثیر شده w_i ها به صورت زیر محاسبه می‌شوند:

$$X_{mis(i)} = 1 \Rightarrow w_i = P(X_{mis(i)} = 1 | Y_i, X_{obs(i)}, \Delta_i=0)$$

$$X_{mis(i)} = 0 \Rightarrow w_i = P(X_{mis(i)} = 0 | Y_i, X_{obs(i)}, \Delta_i=0)$$

۴. مدل رگرسیون لجستیک وزنی را برای حالتی که در آن Δ متغیر وابسته و X_{obs} و X_{mis} و Y متغیر مستقل باشند (نام پارامترهای این مدل را α می‌گذاریم). سپس با استفاده از مجموعه بسط داده شده با n' واحد و وزن‌های w_i برآوردهای به روز شده α ها را به دست می‌آوریم.

۵. با استفاده از برآوردهای جدید α ، مجدداً وزن‌های واحدها را که در مرحله ۳ بیان شد محاسبه کرده و به مرحله ۴ رفته و برآوردهای جدید را با استفاده از وزن‌های جدید محاسبه می‌کنیم. این مراحل را به قدری تکرار می‌کنیم تا به هم‌گرایی برسیم.

پس از انجام عملیات بالا نتایج به دست آمده را بررسی کرده و در صورتی که برآورد پارامتر α_x (پارامتر ضریب متغیر X_{mis}) تقریباً مقداری نزدیک صفر با واریانس بسیار کوچک باشد می‌توان گفت که مجموعه داده دارای سازوکار گم‌شدگی تصادفی می‌باشد.

۳. **گم شدن غیر قابل اغماض^۳ (NI):** در این حالت گم‌شدگی در متغیر کمکی، Y_{mis} ، بستگی به خود Y_{mis} دارد، حتی بعد از اینکه روی سایر متغیرها مشروط شده باشد. برآوردهای هیچ یک از پارامترها از روی داده‌های با مشاهدات کامل، سازگار نخواهند بود و هر گونه استنباطی مستلزم در نظر گرفتن مفروضاتی درباره ارتباط بین متغیرهای Y_{obs} و Y_{mis} و نیز چرایی گم‌شدگی داده‌ها می‌باشد. ذکر این نکته حائز اهمیت است که هیچ تجزیه و تحلیلی از داده‌های دارای گم‌شدگی بدون اغماض، بدون انجام آزمون حساسیت کامل نیست (۱۶، ۱۵).

