
بسم الله الرحمن الرحیم
مرکز اطلاعات و مدارک علمی ایران
شعبه مرکزی

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

داده‌های گمشده، تخصیص و بوتسترپ

پایان نامه کارشناسی ارشد آمار

الهام امیدوار

استاد راهنما

دکتر علی زینل همدانی

دکتر ایوب ساعی

۱۳۸۰

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد رشته آمار خانم الهام امیدوار

تحت عنـــوان

داده‌های گمشده، تخصیص و بوتسترپ

در تاریخ ۸۰/۴/۶ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهائی قرار گرفت.

۱ ـ استاد راهنمای پایان‌نامه	دکتر علی زینل همدانی ـ دکتر ایوب باعی
۲ ـ استاد مشاور پایان‌نامه	دکتر محمد صالحی مرزیجرانی
۳ ـ استاد داور ۱	دکتر مجتبی گنجعلی
۴ ـ استاد داور ۲	دکتر سروش علیمرادی
سرپرست تحصیلات تکمیلی دانشکده	دکتر امیر نادری

تقدیر و تشکر

شکر و سپاس تنها از آن اوست، لکن شکر نعمت‌هایش نیز، حمد اوست پس سپاس می‌گویم آنانکه با نثار جان خویش، آرامش جسم و جان ما را رقم زدند، آنانکه در راه اعتلای حق فداگشتند و آنانکه ایمان سلاحشان بود و عشق مرکبشان.

سپاس می‌گویم آنانی که به گفتهٔ مولایمان، «قد علمنی حرفاً فقد سیرنی عبدا»، مرا به آنچه آموختند بندهٔ خویش ساختند.

تشکر می‌کنم از تمامی اساتید محترمی که در راه یادگیری هر چه بیشتر اینجانب، از هیچ چیز فروگذار نکردند. چه بسا زحمات آنان که فراتر از حس مسؤلیت بوده و چه بساکه در خلال آموزش علم، پرورش روح ایثار شکل گرفته است.

این حقیقت است که محبت‌ها را باید با محبت جبران نمود، امّا زمان لگام گسیخته می‌رود و شرط حضور در مکان فرصت جبران نمی‌دهد و آنچه از من ساخته است جز دعای خیر چیز دیگری نیست.

از استاد بزرگوار جناب آقای دکتر علی همدانی که گاهی اوقات از کثرت زحمات خود، شرمگین می‌شدم، نهایت تشکر را دارم، همچنین از جناب آقای دکتر ساعی که با وجود مشکلاتشان هرگز در برابر مزاحمت‌های من خم به ابرو نیاوردند کمال تشکر را داشته و همواره آرزومند سلامتیشان هستم.

از جناب آقای دکتر صالحی که سمت استاد مشاوری این پایان نامه را تقبل نمودند سپاسگذارم.

از جناب آقای دکتر امیر نادری، سرپرست تحصیلات تکمیلی که همکاریهای لازم را به‌عمل آوردند تشکر می‌کنم.

از آقایان دکتر گنجعلی و دکتر علیمرادی که بعنوان اساتید ممتحن در جلسه دفاعیه شرکت نمودند و بازخوانی رساله را پذیرفتند متشکرم.

از دوستان محترم خانم افسوس و خانم صدرعاملی و خانم انوری که در تایپ پایان‌نامه مرا یاری دادند صمیمانه تشکر می‌کنم و در آخر از تمامی کسانی که به نحوی در تمامی دوران تحصیل از کمک و همراهیشان بهره برده‌ام قدردانی می‌نمایم.

"امید است که خداوند ارتقاء روح و رسیدن به کمال را جبران محبت‌هایشان قرار دهد."

امیدوار

۱۳۸۰

تقدیم به :

سامت مقدس امام عصر(عج).

که عالمی در انتظار آمدنش چشم براهند.

تقدیم به پدر و مادر بزرگوارم

که همواره دلسوزیها و محبت‌هایشان از عمق جان برخاسته است و من هرگز نتوانسته‌ام که حتی ذره‌ای از این دریای بیکران را پاسخ گویم.

تقدیم به همسر مهربان و صبورم

که مشوق ادامه تحصیل و از آن مهمتر، محرک در راه رسیدن به اهدافم بوده است.

و

تقدیم به برادران خوبم

که همیشه این حقیر را به دیدهٔ بزرگ‌بینی نگریسته‌اند.

فهرست

عنوان **صفحه**

فصل دوم: روش بوتسترپ

فصل سوم: داده‌های گمشده، تخصیص و بوتسترپ

چکیده

در بسیاری از مسائل آماری، به دلایل مختلف، تعدادی دادهٔ گمشده وجوددارد که می‌تواند تجزیه و تحلیل اطلاعات را دچار مشکل سازد.

روشهای گوناگون در برخورد با چنین مسائلی پیشنهاد شده است که بعضاً داده‌های گمشده را حذف یا مقادیری را جایگزین می‌کنند.

در مسائلی که برآورد پارامتر مدنظر است، روش بوتسترپ با وجود داده‌های گمشده و محدودیت تعداد نمونه، می‌تواند راه‌گشا باشد. بدین‌منظور در این پایان‌نامه، سه روش اصلی در بوتسترپ همراه با تکنیک‌های بکار گرفته شده در مسئله داده‌های گمشده، مورد توجه قرار گرفته است که عبارتند از :

بوتسترپ ناپارامتری، بوتسترپ با مکانیزم کامل و بوتسترپ با تخصیص چندگانه. در دو روش اول، با استفاده از تعداد تکرار زیاد نمونه‌های بوتسترپ، برآورد فاصله‌ای برای پارامتر مورد علاقه θ حاصل می‌شود و در روش سوم با بکارگیری تخصیص چندگانه بعنوان تکنیکی در جایگزین کردن داده‌های گمشده، روشی محاسباتی و کارا برای محاسبه بازهٔ اطمینان ارائه خواهد شد که با توجه به اختلافات نظری موجود در روش تخصیص چندگانه و بوتسترپ، ارتباط جالب توجهی دیده می‌شود.

نهایتاً نتایج حاصل از هر سه روش، معایب و مزایای تکنیک‌های بکار گرفته شده، مورد بحث و بررسی قرار می‌گیرد.

فصل اول
داده‌های گمشده

۱-۱- مقدمه

بحث داده‌های گمشده، غالباً در بسیاری از مسائل مختلف آماری به چشم می‌خورد که گاهی آماردان برای آنکه خود را با حل اینگونه مسائل درگیر نسازد، سعی می‌کند تا چنین داده‌هایی را کنار بگذارد، در صورتیکه در کتابها و مقالاتی که در این زمینه ارائه گردیده، راه حلهای جالب توجهی دیده می‌شود. داده‌های گمشده به دلایل بسیاری ممکن است به وجود آید. به‌عنوان‌مثال در بررسی‌ها، امکان دارد پاسخ دهنده جواب دادن به پرسشی را رد کند چراکه مثلاً بخواهد آنرا در خلوت و پنهانی پاسخ گوید، یا اینکه شخصی منظور از سئوال را نفهمد یا پاسخ آنرا نداند. گاهی اوقات نیز محدودیت زمانی باعث بی پاسخ ماندن بعضی از سئوالها است. در تمام این موارد، سئوال بی پاسخ، یک داده گمشده محسوب می‌شود.

در تحقیقات نیز زمینه برای ایجاد داده‌های گمشده، زیاد است. بعنوان نمونه، ممکن است محقق گرفتن اندازه‌ای را فراموش کند. مثلاً: نبض بیمار. که این نیز یک داده گمشده به حساب می‌آید.

در برخی موارد، وجود داده‌های گمشده مسائل جدی‌ای را سبب می‌شود و چون تشخیص جدی بودن این داده‌ها کار آسانی نیست، چنین داده‌هایی قابل اطمینان نخواهند بود. در مواردی که نتایج در تصمیم گیریها تأثیر قابل توجهی خواهند داشت، مواجه شدن با داده‌های گمشده حساسیت بیشتری ایجاد می‌کند.

۱-۱-۱- تاریخچه

تاریخچه داده‌های گمشده[۱] بسیار گسترده است. به طور حتم آمار دانان از زمانی که برروی داده‌های آماری تجزیه و تحلیل‌هایی انجام داده‌اند با مسئله گمشدگی داده‌ها مواجه بوده‌اند، زیرا که عملاً امکان عدم مشاهده بعضی از داده‌ها، وجود دارد. آنچه در این بخش بعنوان تاریخچه ارائه می‌گردد، خلاصه‌ای از کارهایی است که نویسندگان مختلف برروی این مبحث انجام داده‌اند.

حدوداً ۷۰ سال پیش ویشارت والان[۲] در سال ۱۹۳۰ [۱] روش برآورد محصول را از یک کرت گمشده[۳]، مورد بحث قرار دادند. مقاله‌ای تحت عنوان "تکنیک‌های کرت گمشده" در سال ۱۹۴۶ [۲] توسط اندرسون[۴] ارائه شد. در اوایل سال ۱۹۵۸ برآورد مقادیر گمشده برای تجزیه و تحلیل داده‌های ناکامل[۵] را ویلکینسون[۶] [۳] مورد مطالعه قراز داد و همین نویسنده در اواخر همین سال تجزیه واریانس را برای داده‌های ناکامل به انجام رسانید [۴].

در سال ۱۹۶۶ الاشف و افیفی[۷] مقاله‌ای تحت عنوان «مشاهدات گمشده در آماره‌های چند متغیره» [۵] ارائه دادند که این مقاله مروری بر تحلیل‌های انجام شده تا آن زمان می‌باشد.

مقاله پریس[۸] در سال ۱۹۷۱ [۶] روشهای تکراری را برای مقادیر گمشده در آزمایش‌ها معرفی می‌کند و در همین سال با همکاری هارتلی و هاکینگ[۹] [۷] تجریه و تحلیل داده‌های ناکامل ارائه می‌گردد.

دمپستر، لیرد و روبین[۱۰] در سال ۱۹۷۷ [۸] ماکزیمم در ستنمایی برای داده‌های ناکامل را ابداع کردند و پس از آن در ۱۹۷۸ روشی تحت عنوان قانون اطلاعات گمشده[۱۱] توسط دو آماردان بنامهای وودبوری و ارچارد[۱۲]

۱- Missing Data
۲- Wishart & Allan
۳- Missing Plot
۴- Anderson
۵- Incomplete data
۶- Wilkinson
۷- Elashoff & Afifi
۸- Preece
۹- Hartley & Hocking
۱۰- Laird & Rubin
۱۱- Missing Information
۱۲- Woodbury & Orchard

[۹] معرفی گردید.

در فواصل هرکدام از این سالها نیز مقالات بسیاری در رابطه با داده‌های ناکامل، مقادیر مشاهده نشده[۱]، داده‌های گمشده و ... موجود است.

در سال ۱۹۸۲ لیتل[۲] [۱۰] داده‌های ناکامل را در بررسیهای نمونه‌ای در مقاله‌ای منتشر ساخت. شاید آنچه تقریباً تمام روشهای مختلف در این‌گونه مسائل را جمع آوری کرده است کتابی است با نام "تجزیه و تحلیل آماری با داده‌های گمشده" از همین نویسنده با همکاری روبین که در سال ۱۹۸۷ [۱۱] انتشار یافت. در این کتاب می‌توان خلاصه‌ای از کارهای انجام شده در رابطه با مسئله داده‌های گمشده را مشاهده نمود.

لازم به ذکر است، از آنجاکه هیچ علمی تاکنون در یک مقطع زمانی مشخص در پیشرفت خود متوقف نشده است پس از سال ۱۹۸۷ نیز تعداد کثیری از مقالات در همین زمینه ارائه گردیده است که از آنجمله می‌توان به مقاله وی و تنر[۳] (۱۹۹۰) [۱۲] اشاره نمود که دو روش مشابه را در مسئله داده‌های گمشده مطرح می‌سازند.

* * * * * *

۱-۲- روشهای سریع در مسأله داده‌های گمشده

در تحلیلهای چند متغیره هر متغیر ممکن است تنها تعداد کمی پاسخ گمشده داشته باشد ولی در ترکیب متغیرها با هم، تعداد داده‌های گمشده زیاد خواهد بود. ساده‌ترین راهی راکه در حل این مشکل پیشنهاد می‌کنند، حذف مواردی است که در آنها دادهٔ گمشده به چشم می‌خورد، [۱۱، ۱۳ و ۱۴]. این عمل، گاهی اوقات سبب می‌شود که تعداد داده‌های باقی مانده برای تجزیه و تحلیل، کفایت نکند و یا آنکه نتایج حاصل معنی‌دار نباشد و یا بعلت آنکه ممکن است مواردی که تجزیه و تحلیل با آنها صورت گرفته، نمونه‌ای تصادفی از کل موارد نتیجه ندهد، نتایج گمراه کننده و غلط انداز شود.

برای مثال، یک تحلیل رگرسیونی برای پیش‌بینی مالکیت خانه براساس سن و سابقهٔ فرهنگی انجام شده

۱- Unobserved Value

۲- Little

۳- Wei & Tanner

است. چنانچه جدول (۱ـ۱) نشان می‌دهد، موارد گمشده با علامت "." نشان داده شده که اگر از مواردی که هر یک از متغیرها پاسخ گمشده دارد، چشم پوشی شود، تعداد کل نمونه‌ها در تحلیل، به طور جدی کاهش می‌یابد.

جدول ۱-۱: نمونه ۸ تایی برای پیش‌بینی مالکیت خانه براساس سن و سابقه فرهنگی

وضعیت شغلی	سابقه فرهنگی	خانه	جنس	سن	نمونه
غیر رسمی	۱۶	خیر	زن	.	۱
غیر رسمی	.	خیر	مرد	۲۲	۲
رسمی	۲۰	.	مرد	۳۹	۳
رسمی	.	بله	زن	.	۴
غیررسمی	۱۶	بله	.	۴۰	۵
.	۱۶	خیر	زن	۲۲	۶
رسمی	۱۸	بله	مرد	۳۵	۷
رسمی	۲۰	بله	مرد	۳۹	۸

در مثالی دیگر، می‌توان گمراه کنندگی نتایج را دید. در جدول (۱ـ۲) و (۱ـ۳) دو بررسی نشان داده می‌شود که در اولی داده گمشده وجود دارد در حالیکه در جدول بعد داده‌ها کاملاً جمع آوری شده‌اند. میانگین سن در جدول (۱ـ۲) برابر با ۲۹ و در جدل (۱ـ۳)، ۳۹ می‌باشد و این تفاوت آشکار نشان می‌دهد که کار کردن با داده‌های گمشده غالباً نتایج گمراه کننده‌ای خواهد داشت.

جدول ۱-۲: نمونه ۸ تایی با دادهٔ گمشده

جنس	سن	نمونه
زن	.	۱
مرد	.	۲
مرد	۳۹	۳
زن	.	۴
مرد	۴۲	۵
زن	.	۶
مرد	۳۷	۷
مرد	۳۹	۸

جدول ۱-۳: نمونهٔ ۸ تایی بدون دادهٔ گمشده

جنس	سن	نمونه
زن	۲۱	۱
مرد	۲۲	۲
مرد	۳۹	۳
زن	۲۰	۴
مرد	۴۲	۵
زن	۱۸	۶
مرد	۳۷	۷
مرد	۳۹	۸

بنابراین این سئوال مطرح می‌شود که تجزیه مشاهداتی راکه داده گمشده دارند، چگونه باید انجام داد؟

روشهای متفاوتی در این رابطه وجود دارد ولی آنچه تحت عنوان روشهای سریع از آن نام برده‌اند، چهار روش ساده است، [۱۱، ۱۳ و ۱۴]. در هر ۴ روش سعی بر آن است که به طریقی مشکل داده‌های گمشده برطرف گردد.

روشها عبارتند از :

۱) حذف کامل[۱]

۲) حذف دوبه دو[۲]

۳) جایگزین کردن (تخصیص) میانگین[۳]

۴) اطلاعات موجود[۴]

۱-۲-۱- حذف کامل

این روش همان روشی است که در دو مثال قبل مورد بحث قرارگرفت. اگر k متغیر موجود باشد در هر نمونه از این k متغیر، حتی اگر تنها یکی از متغیرها بی‌پاسخ مانده، آن نمونه به طورکامل از لیست خارج می‌شود. معایب این روش همانطورکه قبلاً ذکر شد، قابل اطمینان نبودن نتایج حاصل وکاهش زیاد نمونه‌هاست. این روش با حروف اختصاری LD (Listwise Deletion) نشان داده می‌شود.

۱-۲-۲- حذف دو به دو

روش حذف دوبه‌دوکه با حروف اختصاری PD (Pairwise Deletion) نشان داده می‌شود، برای هر دو متغیر، کواریانس را از همه نمونه‌هایی که این دو متغیر با هم مشاهده شده‌اند، برآورد می‌کند.

لازم به ذکر است که تجزیه و تحلیل داده‌های چند متغیره معمولاً با در نظرگرفتن یک مدل چند متغیره، انجام می‌گیرد و در ابتدای کار، برآورد بردار میانگین و ماتریس واریانس ـ کواریانس محاسبه می‌گردد.

به وضوح دیده می‌شود که روش PD اطلاعات بیشتری را از داده‌ها بکار می‌گیرد وکاراتر از روش LD است. اما یکی از معایب آشکار این روش استفاده از اندازه نمونه‌های متفاوت در برآورد هرکواریانس است.

۱ـ Listwise Deletion

۲ـ Pairwise Deletion

۳ـ Mean Imputation

۴ـ Available-Case Method