

تحلیل داده‌های چند متغیره با پاسخ‌های آمیخته گسسته با امکان کم‌شدگی پاسخ‌ها و کاربرد آن در آمارگیری‌های مقطعی و طولی

مجری

سازمان آمار جمهوری اسلامی ایران

زحرا رضایی قزوینی

احسان بهرامی سلمانی

مجتبی کنجلی

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پژوهشگاه آماری

۱۳۸۸



## به نام خداوند جان و خرد

### پیشگفتار

از آنجا که داده‌های حاصل از آمارگیری‌های مقطعی و طولی مراکز آماری، حاوی اطلاعات غنی و ارزشمندی است که نیاز به تحلیل دارد، بنابراین تحلیل آماری مناسب این داده‌ها به منظور تصمیم‌گیری برنامه‌ریزان ضروری است. با توجه به این‌که در آمارگیری‌های مقطعی و طولی، تعداد زیادی از سؤالات پرسشنامه حاوی پاسخ‌های چند متغیره‌ی گسسته مورد علاقه‌ای است که در مقیاس‌های ترتیبی یا اسمی اندازه‌گیری می‌شوند، بنابراین بررسی تأثیر برخی دیگر از متغیرها بر روی این متغیرهای وابسته به طور همزمان از طریق معرفی مدل‌های مقطعی و طولی چند متغیره با پاسخ‌های گسسته الزامی است. از طرفی در بیش‌تر آمارگیری‌های طولی و مقطعی که در مراکز آماری دنیا اجرا می‌شود گریز از داده‌های گم‌شده امری اجتناب‌ناپذیر است. بنابراین معرفی مدل‌هایی که قابلیت بررسی داده‌های گم‌شده و تأثیر آن‌ها بر برآوردها را داشته باشد نیز از اهمیت بسزایی برخوردار است. به همین منظور، پژوهشکده‌ی آمار با توجه به رسالت خود در زمینه‌ی اجرای طرح‌های پژوهشی با هدف تجزیه و تحلیل آمار، اجرای طرح مذکور را در دستور کار قرار داده است. در این رابطه یک گروه مطالعاتی تشکیل شد که گزارش حاضر نتیجه‌ی مطالعات و بررسی‌های این گروه پژوهشی می‌باشد.

در گروه مذکور، سرکار خانم سمانه افتخاری مهابادی به عنوان مجری طرح و سرکار خانم دکتر زهرا رضایی قهرودی و آقای دکتر احسان بهرامی سامانی به عنوان همکار عضویت داشته‌اند که در این‌جا از یکایک این عزیزان تشکر و قدردانی می‌شود.

جناب آقای دکتر مجتبی گنجعلی نیز به عنوان مشاور طرح همکاری نموده‌اند که بدین وسیله از ایشان سپاسگزاری می‌شود. داوری طرح نیز به عهده‌ی جناب آقای دکتر حمید پزشک بوده است که بدین وسیله از راهنمایی‌های صمیمانه‌ی ایشان نیز تشکر و قدردانی می‌شود.

از خوانندگان محترم تقاضا می‌شود، نظرات اصلاحی خود را در ارتباط با محتوای مجموعه‌ی حاضر به گروه پژوهشی طرح‌های فنی و روش‌های آماری منعکس نمایند.

گروه پژوهشی طرح‌های فنی و روش‌های آماری



## مقدمه

در اغلب آمارگیری‌های طولی تعداد زیادی از سئوالات پرسشنامه حاوی متغیرهای گسسته مورد علاقه‌ای می‌باشند که در مقیاس‌های ترتیبی یا اسمی اندازه‌گیری می‌شوند و پژوهشگران علاقه‌مند به بررسی تاثیر دیگر متغیرها به طور همزمان بر این گونه متغیرها هستند. از طرف دیگر در بیشتر آمارگیری‌های طولی و مقطعی که در مراکز آماری دنیا اجرا می‌شود، گریز از داده‌های گم‌شده امری اجتناب‌ناپذیر است. بنابراین معرفی مدل‌هایی که قابلیت بررسی داده‌های گم‌شده و تاثیر آنها بر برآوردها را داشته باشد، ضروری است. در این راستا ابتدا مدل‌های مناسب برای متغیرهای گسسته با مقیاس‌های اسمی و ترتیبی که متغیرهای دودویی را نیز به صورت حالتی خاص شامل می‌شوند، در مطالعات مقطعی، طولی و طولی چند متغیره مورد بازبینی قرار می‌گیرد. سپس مدلی انتقالی با ضرایب تصادفی برای مطالعات طولی چند متغیره با پاسخ‌های آمیخته به منظور بررسی همبستگی بین متغیرهای پاسخ در طول زمان و همچنین پاسخ‌های چند متغیره اندازه‌گیری شده برای هر فرد یا خوشه در یک زمان مشخص، معرفی می‌گردد. مسئله داده‌های گمشده در چنین مطالعات طولی با استفاده از رویکرد بیزی و روش داده‌افزایی برای جانهی مقادیر گمشده مورد بحث و بررسی قرار می‌گیرد.

این طرح پژوهشی شامل چهار فصل می‌باشد. در فصل اول مقدمات و مفاهیم اساسی مورد نیاز در مورد مدل بندی داده‌های گسسته و مفاهیم گم‌شدگی مورد بررسی قرار می‌گیرند. در فصل دوم به مدل بندی مطالعات طولی تک متغیره با پاسخ‌های گسسته با امکان داده‌های گم‌شده پرداخته شده است. همچنین مطالعات مقطعی و طولی چندمتغیره با پاسخ‌های گسسته در فصل سه تشریح و مدل‌های مناسب برای آنها با امکان گم‌شدگی پاسخ‌ها ارائه شده است.

در فصل چهارم سه مجموعه داده مورد تحلیل و مدل بندی قرار می‌گیرد که از میان آن‌ها اولی در مورد داده‌های مربوط به سرپرستان خانوار طرح هزینه و درآمد سال ۱۳۸۶، جمع آوری شده توسط مرکز آمار ایران، می‌باشد. در این داده‌ها به بررسی تاثیر متغیرهای تبیینی موجود بر روی سه پاسخ گسسته وضعیت فقر خانوار، وضعیت درآمد خانوار و وضع فعالیت اقتصادی سرپرست خانوار به طور همزمان، با استفاده از مدل‌های آمیخته معرفی شده در این طرح، پرداخته شده است. کاربرد دوم شامل داده‌های طرح آمارگیری نیروی کار مرکز آمار ایران در فصل پاییز دو سال متوالی ۱۳۸۵ و ۱۳۸۶ برای بررسی عوامل موثر بر تغییر وضع فعالیت اقتصادی اعضای خانوار با در نظر گرفتن همبستگی موجود میان اعضای خانوار می‌باشد. در مثال سوم به تحلیل داده‌های مربوط به طرح آمارگیری پانلی خانوارهای بریتانیا در سه سال متوالی ۲۰۰۳ تا ۲۰۰۵ برای مدل بندی پاسخ ترتیبی رضایت مندی از زندگی همراه با متغیر اسمی وضع فعالیت اقتصادی می‌پردازیم. در این مثال از مدل انتقالی با ضرایب تصادفی برای پاسخ‌های طولی چند متغیره اسمی-ترتیبی استفاده می‌نماییم. همچنین به دلیل امکان گم‌شدگی در پاسخ رضایت مندی از زندگی از روش داده افزایی بی‌زی برای در نظر گرفتن این مسئله استفاده خواهیم نمود.

# فهرست مندرجات

۱	مقدمات و مفاهیم اساسی	۱
۱	مقدمه	۱.۱
۲	انواع مقیاس‌های داده‌های گسسته	۲.۱
۴	مطالعات مقطعی با پاسخ‌های تک متغیره گسسته	۳.۱
۵	پاسخ تک متغیره دودویی	۱.۳.۱
۷	پاسخ تک متغیره اسمی	۲.۳.۱
۹	پاسخ تک متغیره ترتیبی	۳.۳.۱
۱۷	مطالعات همراه با پاسخ گم شده	۴.۱
۱۸	الگوهای مختلف گم شدن	۱.۴.۱

۲۰	مکانیسم‌های مختلف گم‌شدن	۲.۴.۱
۲۲	روش‌های تحلیل داده‌ها با مقادیر گم‌شده	۳.۴.۱
۲۶	روش‌های برآورد پارامترهای مدل	۵.۱
۲۷	روش ماکسیمم درست‌نمایی	۱.۵.۱
۳۱	روش پیزی	۲.۵.۱
۴۳	<b>۲ مطالعات طولی تک متغیره با پاسخ‌های گسسته</b>	
۴۳	مقدمه	۱.۲
۴۴	مطالعات طولی	۲.۲
۴۵	روش‌های مدل‌بندی داده‌های طولی	۳.۲
۴۶	مدل‌های حاشیه‌ای	۱.۳.۲
۴۹	مدل‌های انتقالی زنجیر مارکف	۲.۳.۲
۵۳	مدل ضرایب تصادفی	۳.۳.۲
۵۸	مدل‌های طولی برای پاسخ تک متغیره گسسته با امکان گم‌شدگی	۴.۲



۶۳	۳	مطالعات مقطعی و طولی چند متغیره با پاسخ‌های گسسته
۶۳	۱.۳	مقدمه
۶۴	۲.۳	مطالعات مقطعی چند متغیره
	۱.۲.۳	مدل دارای ضرایب تصادفی برای پاسخ‌های چند متغیره
	۶۴	اسمی مقطعی
	۲.۲.۳	مدل دارای ضرایب تصادفی برای پاسخ‌های چند متغیره
	۶۵	ترتیبی مقطعی
	۳.۲.۳	مدل دارای ضرایب تصادفی برای پاسخ‌های چند متغیره
	۶۶	اسمی - ترتیبی مقطعی
۷۰	۳.۳	مطالعات طولی چند متغیره
	۱.۳.۳	مدل انتقالی دارای ضرایب تصادفی برای پاسخ‌های طولی
	۷۰	چند متغیره گسسته با مقیاس یکسان
	۴.۳	مدل انتقالی با ضرایب تصادفی برای پاسخ‌های طولی چند متغیره
۷۴		اسمی - ترتیبی
۸۰	۴	مثال‌های کاربردی در آمارگیری‌های مقطعی و طولی

۸۰	مقدمه	۱.۴
۸۲	مثال ۱: داده‌های طرح هزینه و درآمد خانوار سال ۱۳۸۶	۲.۴
۸۶	معرفی مدل	۱.۲.۴
۹۰	نتایج	۲.۲.۴
۹۷	مثال ۲: داده‌های طرح نیروی کار سال‌های ۱۳۸۵ و ۱۳۸۶	۳.۴
۱۰۰	معرفی مدل	۱.۳.۴
۱۰۲	نتایج	۲.۳.۴
۱۱۱	مثال ۳: داده‌های طرح آمارگیری پانلی خانوارهای بریتانیا	۴.۴
۱۱۲	معرفی مدل	۱.۴.۴
۱۱۸	نتایج	۲.۴.۴
۱۲۲	A برنامه‌های محاسباتی مربوط به مثال‌های کاربردی	
۱۵۲	B واژه‌نامه‌ی فارسی به انگلیسی	
۱۵۶	مراجع	

## فصل ۱

# مقدمات و مفاهیم اساسی

### ۱.۱ مقدمه

از آنجا که طرح حاضر به دنبال بررسی مدل‌های مناسب برای تحلیل داده‌های شامل پاسخ‌های گسسته با امکان گم‌شدگی می‌باشد، در این فصل به معرفی مفاهیم اساسی پاسخ‌های گسسته و داده‌های گم‌شده در مطالعات مقطعی می‌پردازیم. در این راستا در بخش ۲، انواع مقیاس بندی داده‌های گسسته و لزوم ورود متغیرها با مقیاس واقعی خود در مرحله تحلیل داده‌ها را مورد بررسی قرار می‌دهیم. در بخش ۳، مروری بر مطالعات مقطعی با حضور پاسخ‌های گسسته برای انواع مقیاس‌های دودویی، اسمی و ترتیبی و مدل‌های مناسب برای اینگونه پاسخ‌ها را خواهیم داشت. مسئله گم‌شدن و مفاهیم مورد نیاز در برخورد با این مسئله نیز در بخش ۴ به طور مختصر مورد توجه قرار خواهد گرفت. در انتها، برخی روش‌های برآورد پارامترهای مدل همچون ماکسیمم درست‌نمایی و روش‌های بیزی و همچنین روش‌های شبیه‌سازی مورد نیاز در برآورد مدل‌های بیزی در بخش ۵ ارائه می‌گردد.

## ۲.۱ انواع مقیاس‌های داده‌های گسسته

کاملاً روشن است که نوع داده همچون نوع مسئله‌ای که آماردان با آن مواجه است با تغییر زمینه تحقیق، تغییر بسیاری می‌یابد. در نتیجه روش‌هایی که در زمینه‌ای خاص مفید می‌باشند ممکن است کاربرد اندکی را در دیگر زمینه‌ها داشته باشند. برای مثال، در علوم فیزیکی داده‌ها لزوماً کمی با مقیاس‌های دلخواه هستند. در حالی که در علوم اجتماعی و علوم زیستی، داده‌های کیفی بسیار رایج است. این اندازه‌گیری‌های کیفی معمولاً مقادیری را در مجموعه‌ای محدود از رده‌ها اختیار می‌نمایند که ممکن است ترتیبی یا دارای مقیاس اسمی باشند.

استیونز<sup>۱</sup> (۱۹۵۱، ۱۹۵۹ و ۱۹۶۸) به بررسی انواع مقیاس‌ها و ارتباط آنها با روش‌های آماری مورد استفاده پرداخته و چهار مقیاس اسمی، ترتیبی، فاصله‌ای و نسبتی را معرفی نمود. به طور خاص، متغیرهای رده بندی شده دارای دو نوع مقیاس اسمی و ترتیبی می‌باشند.

- متغیرهای اسمی: متغیرهای دارای رده‌های بدون ترتیب طبیعی، "اسمی" نامیده می‌شوند. مثال‌هایی از متغیرهای اسمی عبارتند از: وضع فعالیت اقتصادی (شامل رده‌های شاغل، بیکار و غیر فعال)، دین یا مذهب (شامل رده‌های مسلمان، کاتولیک، پروتستان، یهودی و غیره)، شیوه حمل و نقل برای رفتن به محل کار (سواری، دوچرخه، اتوبوس، مترو و پیاده) و محل سکونت (آپارتمان، مجتمع، خانه و غیره). برای متغیرهای اسمی، ترتیب لیست شدن رده‌ها اهمیتی ندارد.

- متغیرهای ترتیبی: بسیاری از متغیرهای گسسته یا رده بندی شده دارای رده‌های مرتب

<sup>۱</sup> Stevnes

هستند. چنین متغیرهایی "ترتیبی" نامیده می‌شوند. مثال‌هایی از این نوع متغیرها عبارتند از: وضعیت فقر (شبه فقیر، غیر فقیر، شبه غنی و غنی) که بر اساس شاخص فقر به دست می‌آیند، فلسفه سیاسی (آزادی خواه، میانه رو، محافظه کار) و شرایط بیمار (خوب، نسبتاً خوب، وخیم، بحرانی). در متغیرهای ترتیبی، ترتیب میان رده‌ها از اهمیت برخوردار است اما فاصله میان رده‌ها نامعلوم است. مثلاً فردی که به عنوان میانه رو در نظر گرفته شود بیشتر از یک محافظه کار، آزادی خواه است اما هیچ مقدار عددی نمی‌تواند میزان این تفاوت را بیان کند. در روش‌های تحلیل داده‌های ترتیبی از ترتیب رده‌ها به عنوان یک اطلاع در مدل بندی استفاده می‌نمایند.

- متغیرهای فاصله‌ای: متغیرهای فاصله‌ای و نسبتی، متغیرهایی هستند که به جز خاصیت ترتیبی دارای فواصل عددی میان مقادیرشان هستند همچون فشارخون و درآمد سالانه. گاهی یک متغیر فاصله‌ای، نسبتی نیز نامیده می‌شود در صورتی که نسبت مقادیر نیز معتبر و قابل تفسیر باشد. در مقیاس نسبتی صفر تعریف شده وجود دارد ولی در مقیاس فاصله‌ای صفر تعریف شده نداریم به عنوان مثال، درجه حرارت دارای مقیاس فاصله‌ای و وزن در مقیاس نسبتی اندازه گیری می‌شود.

مقیاس اندازه گیری یک متغیر، تعیین کننده روش‌های آماری مناسب برای تحلیل آن متغیر است. در مراتب اندازه گیری، متغیرهای فاصله‌ای بالاترین، متغیرهای ترتیبی بعد از آنها و متغیرهای اسمی پایین ترین رتبه را دارند (اگرستی<sup>۲</sup>، ۲۰۰۲). روش‌های آماری مناسب برای یک نوع از مقیاس، قابل به کارگیری برای متغیرهای در مراتب بالاتر و نه برای مراتب پایین تر است. برای نمونه، روش‌های آماری برای متغیرهای اسمی را می‌توان برای متغیرهای

<sup>۲</sup> Agresti

ترتیبی با صرف نظر از ترتیب موجود استفاده کرد اما روش‌های مناسب برای متغیرهای ترتیبی را نمی‌توان برای متغیرهای اسمی به کار برد زیرا هیچگونه ترتیب معناداری میان رده‌ها موجود نیست. پس بهترین راهبرد به کارگیری روش آماری متناظر با مقیاس واقعی هر متغیر است. از آنجا که طرح حاضر در راستای مواجهه با متغیرهای گسسته یا رده بندی شده است، به تحلیل متغیرهای دودویی، اسمی و ترتیبی خواهیم پرداخت. این روش‌ها قابل به کارگیری برای متغیرهای فاصله‌ای با تعداد مقادیر مجزای کم (همچون تعداد افراد شاغل خانوار، تعداد دفعات ازدواج و تعداد فرزندان) یا برای متغیرهای فاصله‌ای که مقادیرشان به صورت رده‌هایی ترتیبی طبقه‌بندی شده‌اند (همچون مدت زمان بیکاری با رده‌های کمتر از ۳ ماه، ۳ تا ۶ ماه، ۶ تا ۱۲ ماه، ۱۲ تا ۲۴ ماه و بیشتر از ۲۴ ماه) نیز می‌باشند.

### ۳.۱ مطالعات مقطعی با پاسخ‌های تک متغیره گسسته

در این بخش به معرفی و بررسی مطالعات مقطعی با پاسخ تک متغیره گسسته می‌پردازیم. مطالعات مقطعی تک متغیره، آن دسته از مطالعاتی هستند که در آنها تنها یک متغیر پاسخ در مقطعی معلوم از زمان مورد بررسی قرار می‌گیرد و می‌توان آن را به وسیله یک بردار از متغیرهای کمکی  $\underline{x} = (x_1, \dots, x_m)'$  شرح و تعبیر کرد. در این مطالعات متغیر پاسخ می‌تواند پیوسته، گسسته دودویی، اسمی و یا ترتیبی باشد و همچنین متغیرهای کمکی می‌توانند پیوسته یا از نوع رده بندی شده باشند. با توجه به اینکه هدف این طرح تحلیل پاسخ‌های گسسته است، در زیر بخش‌های آتی به ترتیب به بررسی مدل‌های مناسب برای انواع متغیرهای پاسخ دودویی، اسمی و ترتیبی می‌پردازیم.

## ۱.۳.۱ پاسخ تک متغیره دودویی

فرض کنید  $Y_i$  نشان دهنده متغیر پاسخ دودویی مورد علاقه برای فرد  $i$  ام باشد  $(i = 1, \dots, n)$ . بنابراین هر مشاهده دو برآمد ممکن دارد که با ۰ و ۱ نمایش داده می شود. توزیع مناسب برای چنین پاسخی، برنولی با احتمال پیروزی  $p_i = P(Y_i = 1 | x_{1i}, \dots, x_{mi})$  برای فرد  $i$  ام می باشد که پاسخ ها برای افراد مختلف، همچون مدل های رگرسیونی، مستقل فرض می گردند. یعنی:

$$Y_i \sim \text{Bernolli}(p_i)$$

حال با به کارگیری مفهوم رگرسیون که همان امید ریاضی شرطی متغیر پاسخ به شرط متغیرهای کمکی می باشد، می بایست به مدل بندی تابعی از

$$p_i = E(Y_i | x_{1i}, \dots, x_{mi}) = P(Y_i = 1 | x_{1i}, \dots, x_{mi})$$

به عنوان ترکیبی خطی از متغیرهای کمکی به صورت زیر پرداخت:

$$g(p(Y_i = 1 | x_{1i}, \dots, x_{mi})) = \alpha + \sum_{k=1}^m \beta_k x_{ki} \quad (1.1)$$

که در آن  $g$  تابع ربط نیز نامیده می شود. اما باید توجه داشت که استفاده از تابع ربط همانی همچون مدل های رگرسیون خطی قابل قبول نیست زیرا این تابع ربط باعث می شود سمت چپ رابطه (۱.۱) همواره عددی میان صفر و یک باشد در حالی که سمت راست معادله می تواند هر عدد حقیقی را اختیار نماید. گزینه های مختلفی برای انتخاب تابع ربط وجود دارد که مشهورترین آنها تابع ربط لوجیت  $[Logit(a) = Log\{a/(1-a)\}]$  می باشد. در رگرسیون