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

مجری

سمانه افتخاری مهلبادی

زهرا رضایی قهرودی

احسان بهرامی سامانی

مجتبی گنجعلی

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پژوهشکده‌ی آمار

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 اجتناب ناپذیر 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 گسستهٔ مورد علاقه‌ای می‌باشند که در مقیاس‌های ترتیبی یا اسمی اندازه گیری می‌شوند و پژوهشگران علاقه مند 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 اجتناب ناپذیر 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 اندازه گیری 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 داده افزایی 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 گسستهٔ وضعیت فقر خانوار، وضعیت درآمد خانوار و وضع فعالیت اقتصادی سرپرست خانوار به طور همزمان، با استفاده از مدل‌های آمیختهٔ 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 مدل‌بندی پاسخ ترتیبی رضایت‌مندی از زندگی همراه با متغیر اسمی وضع فعالیت اقتصادی می‌پردازیم. در این مثال از مدل انتقالی با ضرایب تصادفی برای پاسخ‌های طولی چند متغیرهٔ اسمی—ترتیبی استفاده می‌نماییم. همچنین به دلیل امکان گم‌شدگی در پاسخ رضایت‌مندی از زندگی از روش داده افزایی بیزی برای در نظر گرفتن این مسئله استفاده خواهیم نمود.

فهرست مندرجات

۱ مقدمات و مفاهیم اساسی ۱

۱.۱ مقدمه ۱

۲.۱ انواع مقیاس‌های داده‌های گسسته ۲

۳.۱ مطالعات مقطعی با پاسخ‌های تک متغیرهٔ گسسته ۴

۱.۳.۱ پاسخ تک متغیرهٔ دودویی ۵

۲.۳.۱ پاسخ تک متغیرهٔ اسمی ۷

۳.۳.۱ پاسخ تک متغیرهٔ ترتیبی ۹

۴.۱ مطالعات همراه با پاسخ گم‌شده ۱۷

۱.۴.۱ الگوهای مختلف گم‌شدن ۱۸

۲.۴.۱ مکانیسم‌های مختلف گم‌شدن ۲۰
۳.۴.۱ روش‌های تحلیل داده‌ها با مقادیر گم‌شده ۲۲

۵.۱ روش‌های برآورد پارامترهای مدل ۲۶
۱.۵.۱ روش ماکسیمم درستنمایی ۲۷
۲.۵.۱ روش بیزی ۳۱

۲ مطالعات طولی تک متغیره با پاسخ‌های گسسته ۴۳

۱.۲ مقدمه ۴۳

۲.۲ مطالعات طولی ۴۴

۳.۲ روش‌های مدل‌بندی داده‌های طولی ۴۵
۱.۳.۲ مدل‌های حاشیه‌ای ۴۶
۲.۳.۲ مدل های انتقالی زنجیر مارکف ۴۹
۳.۳.۲ مدل ضرایب تصادفی ۵۳

۴.۲ مدل‌های طولی برای پاسخ تک متغیرهٔ گسسته با امکان گم‌شدگی ۵۸

۳ مطالعات مقطعی و طولی چند متغیره با پاسخ‌های گسسته ۶۳

۱.۳ مقدمه ۶۳

۲.۳ مطالعات مقطعی چند متغیره ۶۴

۱.۲.۳ مدل دارای ضرایب تصادفی برای پاسخ‌های چند متغیرهٔ اسمیِ مقطعی ۶۴

۲.۲.۳ مدل دارای ضرایب تصادفی برای پاسخ‌های چند متغیرهٔ ترتیبیِ مقطعی ۶۵

۳.۲.۳ مدل دارای ضرایب تصادفی برای پاسخ‌های چند متغیرهٔ اسمی– ترتیبی مقطعی ۶۶

۳.۳ مطالعات طولی چند متغیره ۷۰

۱.۳.۳ مدل انتقالی دارای ضرایب تصادفی برای پاسخ‌های طولی چند متغیرهٔ گسسته با مقیاس یکسان ۷۰

۴.۳ مدل انتقالی با ضرایب تصادفی برای پاسخ‌های طولی چند متغیرهٔ اسمی– ترتیبی ۷۴

۴ مثال‌های کاربردی در آمارگیری‌های مقطعی و طولی ۸۰

۱.۴ مقدمه ۸۰

۲.۴ مثال ۱: داده‌های طرح هزینه و درآمد خانوار سال ۱۳۸۶ . . . . . ۸۲
۱.۲.۴ معرفی مدل ۸۶
۲.۲.۴ نتایج ۹۰

۳.۴ مثال ۲: داده‌های طرح نیروی کار سال‌های ۱۳۸۵ و ۱۳۸۶ . . . . . ۹۷
۱.۳.۴ معرفی مدل ۱۰۰
۲.۳.۴ نتایج ۱۰۲

۴.۴ مثال ۳: داده‌های طرح آمارگیری پانلی خانوارهای بریتانیا ۱۱۱
۱.۴.۴ معرفی مدل ۱۱۲
۲.۴.۴ نتایج ۱۱۸

A برنامه‌های محاسباتی مربوط به مثال‌های کاربردی ۱۲۲

B واژه‌نامه‌ی فارسی به انگلیسی ۱۵۲

مراجع ۱۵۶

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 مرحلهٔ 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. مسئلهٔ گم‌شدن و مفاهیم مورد نیاز در برخورد با این مسئله نیز دربخش 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 زمینهٔ 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 [۱] (۱۹۵۱، ۱۹۵۹و 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)، شیوهٔ 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

[۱] Stevenes

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، فلسفهٔ سیاسی (آزادی خواه، میانه رو، محافظه کار) و شرایط بیمار (خوب، نسبتا خوب، وخیم، بحرانی). در متغیرهای ترتیبی، ترتیب میان رده‌ها از اهمیت برخوردار است اما فاصلهٔ میان رده‌ها نامعلوم است. مثلا 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، درجهٔ حرارت دارای مقیاس فاصله‌ای و وزن در مقیاس نسبتی اندازه گیری می‌شود.

مقیاس اندازه گیری یک متغیر، تعیین کنندهٔ روش‌های آماری مناسب برای تحلیل آن متغیر است. در مراتب اندازه گیری، متغیرهای فاصله‌ای بالاترین، متغیرهای ترتیبی بعد ازآنها و متغیرهای اسمی پایین ترین رتبه را دارند (اگرستی[۲]، ۲۰۰۲). روش‌های آماری مناسب برای یک نوع از مقیاس، قابل به کار گیری برای متغیرهای در مراتب بالاتر و نه برای مراتب پایین تر است. برای نمونه، روش‌های آماری برای متغیرهای اسمی را می‌توان برای متغیرهای

[۲] 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 ٦ ماه، ٦ تا ۱۲ ماه، ۱۲ تا ۲۴ ماه و بیشتر از ۲۴ ماه) نیز می‌باشند.

۳.۱ مطالعات مقطعی با پاسخ‌های تک متغیرهٔ گسسته

در این بخش به معرفی و بررسی مطالعات مقطعی با پاسخ تک متغیرهٔ 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 وسیلهٔ یک بردار از متغیرهای کمکی $\underline{x} = (x_1, ..., x_m)'$ شرح و تعبیر کرد. در این مطالعات متغیر پاسخ می‌تواند پیوسته، گسستهٔ دودویی، اسمی و یا ترتیبی باشد و همچنین متغیرهای کمکی می توانند 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 متغیرهٔ دودویی

فرض کنید Y_i نشان دهندهٔ متغیر پاسخ دودویی مورد علاقه برای فرد i ام باشد $(i = ۱, \ldots, n)$. بنابراین هر مشاهده دو برآمد ممکن دارد که با ۰ و ۱ نمایش داده می‌شود. توزیع مناسب برای چنین پاسخی، برنولی با احتمال پیروزی $p_i = P(Y_i = ۱|x_{۱i}, \ldots, x_{mi})$ برای فرد i ام می‌باشد که پاسخ‌ها برای افراد مختلف، همچون مدل‌های رگرسیونی، مستقل فرض می‌گردند. یعنی:

$$Y_i \sim Bernolli(p_i)$$

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

$$p_i = E(Y_i|x_{۱i}, \ldots, x_{mi}) = P(Y_i = ۱|x_{۱i}, \ldots, x_{mi})$$

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

$$g(p(Y_i = ۱|x_{۱i}, \ldots, x_{mi})) = \alpha + \sum_{k=۱}^{m} \beta_k x_{ki} \qquad (۱.۱)$$

که در آن g تابع ربط نیز نامیده می‌شود. اما باید توجه داشت که استفاده از تابع ربط همانی همچون مدل‌های رگرسیون خطی قابل قبول نیست زیرا این تابع ربط باعث می‌شود سمت چپ رابطهٔ (۱.۱) همواره عددی میان صفر و یک باشد در حالی که سمت راست معادله می‌تواند هر عدد حقیقی را اختیار نماید. گزینه‌های مختلفی برای انتخاب تابع ربط وجود دارد که مشهورترین آنها تابع ربط لوجیت $[Logit(a) = Log\{a/(۱ - a)\}]$ می‌باشد. در رگرسیون

لوژستیک فرض می‌کنیم:

$$Log\{\frac{p_i}{۱-p_i}\} = \alpha + \sum_{k=۱}^{m} \beta_k x_{ki}$$

که به عبارت سمت چپ، لگاریتم بخت[۳] نیز گفته می‌شود. از دیگر توابع ربط معروف می‌توان به معکوس تابع توزیع تجمعی نرمال اشاره کرد که منجر به مدل پروبیت به فرم زیر می‌شود:

$$p_i = \Phi(\alpha + \sum_{k=۱}^{m} \beta_k x_{ki})$$

همچنین مدل لگ–لگ مکمل[۴] نیز با تابع لگ–لگ $[Log\{-Log(۱-a)\}]$ قابل حصول است:

$$Log\{-Log(۱-p_i)\} = \alpha + \sum_{k=۱}^{m} \beta_k x_{ki}$$

حال با استفاده از توضیحات بالا، تابع درستنمایی مدل با فرض استقلال میان مشاهدات هر واحد با واحد دیگر به فرم زیر خواهد بود:

$$L(\alpha, \beta_۱, \ldots, \beta_m | \underline{y}, \underline{x}) = \prod_{i=۱}^{n} P(Y_i = y_i | \underline{x}, \alpha, \beta_۱, \ldots, \beta_m) = \prod_{i=۱}^{n} p_i^{y_i} (۱-p_i)^{۱-y_i}.$$

که با در نظر گرفتن هر یک از مدل‌های لوژستیک، پروبیت یا لگ–لگ مکمل، تابع درستنمایی با جایگذاری p_i مناسب از رابطهٔ مدل، قابل بیان است.

البته رویکرد احتمالی نیز در راستای دستیابی به انواع مدل‌ها برای پاسخ‌های دودویی با معرفی متغیرهای پنهان قابل استدلال می‌باشد که ما در بررسی پاسخ‌های ترتیبی به آن اشاره خواهیم نمود. مدل‌های معرفی شده در این زیربخش برای پاسخ‌های دودویی قابل اجرا توسط نرم افزارهای موجود همچون نرم افزار R (تابع $glm(.)$) می‌باشند.

[۳] Odds

[۴] Complementry Log-Log

۲.۳.۱ پاسخ تک متغیرهٔ اسمی

فرض کنید Y_i دارای J ردهٔ اسمی باشد. توزیع مناسب برای چنین پاسخی، چندجمله‌ای (J جمله‌ای) با احتمالات $p_{ij} = P(Y_i = j|x_{1i}, \ldots, x_{mi})$ برای $j = 1, \ldots, J$ می‌باشد که پاسخ‌ها برای افراد مختلف مستقل فرض می‌گردند. یعنی:

$$Y_i \sim Multinomial(n_i, p_{i1}, p_{i2}, \ldots, p_{iJ})$$

برای مدل‌بندی چنین پاسخی با استفاده از ایدهٔ مدل‌بندی متغیرهای دودویی در بخش قبلی می‌توانیم هر دو ردهٔ ممکن را همچون متغیرهای دودویی به صورت مدل لوژستیک مدل‌بندی کنیم و در واقع از مدل‌های لوجیت چندجمله‌ای که به طور همزمان لگاریتم نسبت احتمالات برای تمامی $\begin{pmatrix} J \\ 2 \end{pmatrix}$ زوج ممکن از رده‌ها را مدل می‌کند، استفاده کنیم. به هر حال همان طور که در زیر نشان داده خواهد شد، باید توجه داشت که با انتخاب تنها $J - 1$ تا از این زوج‌ها، بقیه اضافی و قابل محاسبه می‌باشند.

• مدل‌های لوجیت چند جمله‌ای

با فرض توزیع چندجمله‌ای برای Y_i با احتمالات $(p_{i1}, p_{i2}, \ldots, p_{iJ})$، مدل‌های لوجیت هر رده از متغیر پاسخ را با یک ردهٔ مبنا که اغلب اولین یا شایع‌ترین رده می‌باشد، مقایسه می‌کنند (اگرستی، ۲۰۰۲ ص ۲۶۸). مدل زیر

$$Log\frac{p_{ij}}{p_{iJ}} = \alpha_j + \sum_{k=1}^{m} \beta_{kj} x_{ki}, \qquad j = 1, \ldots, J-1 \tag{۲.۱}$$

به طور همزمان اثر متغیرهای کمکی $(x_1, \ldots, x_m)$ را بر این $J - 1$ لوجیت تبیین می‌نماید. در مدل (۲.۴) پارامترها وابسته به اندیس j می‌باشند که بیانگر تغییر تاثیرات برحسب ردهٔ پاسخ

مورد مقایسه با ردهٔ مبنا است. به عبارت دیگر این $(J-1)$ معادله، پارامترهای لوجیت برای دیگر جفت‌ها از رده‌های پاسخ را تعیین می‌کنند زیرا برای مثال برای رده‌های a و b داریم:

$$Log\frac{p_{ia}}{p_{ib}} = Log\frac{p_{ia}}{p_{iJ}} - Log\frac{p_{ib}}{p_{iJ}}.$$

حال با توجه به اینکه بر اساس رابطهٔ (۲.۴) به ازای $j = 1, \ldots, J-1$، $p_{ij} = p_{iJ} \times e^{\alpha_j + \sum_{k=1}^{m} \beta_{kj} x_{ki}}$ و $\sum_{j=1}^{J} p_{ij} = 1$ داریم:

$$\sum_{j=1}^{J} p_{ij} = p_{iJ} \times \sum_{j=1}^{J-1} e^{\alpha_j + \sum_{k=1}^{m} \beta_{kj} x_{ki}} + p_{iJ} = 1$$

در نتیجه برای ردهٔ آخر،

$$p_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\alpha_j + \sum_{k=1}^{m} \beta_{kj} x_{ki}}}$$

و برای $j = 1, \ldots, J-1$ داریم:

$$p_{ij} = \frac{e^{\alpha_j + \sum_{k=1}^{m} \beta_{kj} x_{ki}}}{1 + \sum_{j=1}^{J-1} e^{\alpha_j + \sum_{k=1}^{m} \beta_{kj} x_{ki}}}$$

که همان احتمالات نقطه‌ای مورد نیاز در تابع درستنمایی مدل هستند. حال با در نظر گرفتن n مشاهدهٔ مستقل از متغیر اسمی مورد بحث (یعنی برای هر i، $n_i = 1$)، تابع درستنمایی به صورت زیر است:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m | \underline{y}, \underline{x}) = \prod_{i=1}^{n} [P(Y_i = y_i | \underline{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)] = \prod_{i=1}^{n} [\prod_{j=1}^{J} p_{ij}^{y_{ij}}].$$

که در آن y_{ij} برای $i = 1, \ldots, n$ و $j = 1, \ldots, J$ متغیری دودویی است که اگر $y_i = j$ باشد مقدار یک و در غیر این صورت مقدار صفر را اختیار می‌نماید. مدل‌های لوجیت چندجمله‌ای برای پاسخ‌های اسمی را می‌توان با استفاده از تابع $multinom(.)$ در نرم افزار R برآورد نمود.

۳.۳.۱ پاسخ تک متغیرهٔ ترتیبی

اغلب متغیرهای پاسخی که بیش از دو سطح دارند، ترتیبی می‌باشند و در نظر نگرفتن این ترتیب از دست دادن قسمتی از اطلاعات نمونه را در پی خواهد داشت و این مسئله در نمونه‌های با حجم کمتر، مضرتر خواهد بود (اگرستی، ۲۰۰۲، فصل ۱). در این بخش فرض خواهیم کرد که متغیر پاسخ ترتیبی بوده و مقادیر ممکنهٔ آن $1, 2, \ldots, J$ باشد.

به طور مثال متغیر پاسخ می‌تواند وضعیت فقر خانوار باشد که شامل سطوح فقیر، شبه فقیر، غیر فقیر و شبه غنی یا غنی است. این متغیر را می‌توان به صورت ترتیبی در نظر گرفت و به طور خاص تاثیر متغیرهای کمکی مانند وضع تحصیلات سرپرست خانوار، وضعیت مسکن خانوار، بُعد خانوار و وضعیت تاهل سرپرست خانوار بر روی آن می‌تواند مورد بررسی قرار گیرد. مثال دیگر انتظار شغلی افراد بعد از فراغت از تحصیل است که این متغیر می‌تواند شامل سطوح فرصت‌های شغلی ناکافی، تا حدودی کافی و وجود شغل فوراً بعد از فراغت از تحصیل باشد.

متغیر ترتیبی ممکن است از مکانیسم‌های مختلفی ایجاد شده باشد. این متغیر ممکن است از عملیات گروه‌بندی یک متغیر پیوسته ایجاد شده باشد، به طور نمونه در مثال وضعیت فقر خانوار، پاسخ ترتیبی به دست آمده می‌تواند حاصل از گروه بندی متغیر هزینهٔ کل خانوار بر اساس شاخص فقر باشد و در نتیجه پاسخ ترتیبی از مکانیسم گروه‌بندی یک متغیر پیوسته حاصل شود. گروه دیگر متغیرهای ترتیبی وقتی به وجود می‌آیند که پاسخ دهنده با تحلیل مقدار نامعلومی از اطلاعات، قضاوتی را در مورد رتبهٔ مقیاس ترتیبی در نظر بگیرد، مثلاً در مطالعهٔ رضایت شغلی ، هر فرد ممکن است بر اساس خصوصیات خود و کشورش نظری را در مورد میزان رضایتمندی خود داشته باشد.

در مدل‌بندی متغیرهای ترتیبی از مفهوم بخت‌ها به‌جای احتمال استفاده می‌گردد. لازم به ذکر است که یک بخت به سادگی از نسبت دو احتمال به دست می‌آید. ۴ نوع متفاوت بخت (اگرستی ۲۰۰۲) وابسته به نوع مکانیسم تولید متغیر ترتیبی، وجود دارد که در زیر به بررسی انواع این بخت‌ها می‌پردازیم:

الف) بخت‌های تجمعی

بخت‌های تجمعی با نماد $O(cum) = P(Y \leq j)/P(Y \geq j+1)$ نشان داده می‌شود (نماد O برگرفته از نسبت بخت است) که اغلب زمانی که متغیر ترتیبی حاصل از رده‌بندی یک متغیر پیوسته باشد، به کار می‌رود. مانند متغیر نشانگر وضعیت فقر خانوار که از رده‌بندی متغیر پیوستهٔ هزینه کل خانوار حاصل می‌شود.

ب) بخت‌های رده مجاور

بخت‌های رده مجاور که با نماد $O(adj) = P(Y = j)/P(Y = j+1)$ نمایش داده می‌شود، معمولاً هنگامی که احتمالات هر ردهٔ پاسخ مورد علاقه باشد، یعنی زمانی که متغیر ترتیبی، رده‌بندی شدهٔ ذاتی باشد، استفاده می‌شود، مانند متغیر میزان رضایت شغلی افراد.

ج) بخت‌های نسبتی—تسلسلی

بخت‌های تسلسلی دارای دو نوع $O(con1) = P(Y = j)/P(Y \geq j+1)$ و $O(con2) = P(Y \leq j)/P(Y = j+1)$ می‌باشد که مربوط به فرآیندهای تصمیم‌گیری دنباله‌ای است که در آن، گزینه‌ها یا از کم به زیاد (نوع اول) یا از زیاد به کم (نوع دوم) سنجیده می‌شوند. مانند بررسی وضعیت توماری سرطانی که از کوچک تا خیلی بزرگ قابل رده بندی است و در هر مرحله این اندازه به طور دنباله‌ای تغییر وضعیت می‌دهد.

• مدل‌های تجمعی (روش آستانه‌ای)

رایج‌ترین مدل مورد استفاده در رگرسیون ترتیبی بر اساس روش آستانه‌ای می‌باشد. مدل‌های آستانه‌ای متغیر پاسخ مشاهده شده برای واحد i ام، Y_i برای $i = ۱, \ldots, n$ را صورت رده‌بندی شدهٔ یک متغیر پنهان پیوستهٔ U_i در نظر می‌گیرند. اگر متغیر ترتیبی حاصل از مکانیسم گروه‌بندی باشد، می‌توان U_i را متغیر مورد استفاده در رده‌بندی در نظر گرفت و در حالت وجود مکانیسم قضاوت شخصی بر اساس میزان اطلاعات هر فرد، U_i را می‌توان متغیر پنهان در مقیاس پیوسته در نظر گرفت. در هر دو مکانیسم متغیر پنهان برای ایجاد مدل‌های تجمعی استفاده می‌گردد.

برای بردار متغیرهای کمکی مربوط به واحد i ام، $\underline{x_i} = (x_{i۱}, \ldots, x_{im})'$، روش آستانه‌ای، رابطهٔ متغیر قابل مشاهدهٔ $Y_i \in \{۱, ..., J\}$ با متغیر پنهان U_i را برای سطوح مختلف پاسخ $j = ۱, ..., J$ به صورت زیر در نظر می گیرد:

$$Y_i = j \quad \Leftrightarrow \quad \theta_{j-۱} < U_i \leq \theta_j$$

که در آن $-\infty = \theta_۰ < \theta_۱ < ... < \theta_J = \infty$، بدین معنا که Y_i حالت رده‌بندی شدهٔ U_i توسط نقاط آستانه‌ای $\theta_۱, ..., \theta_{J-۱}$ می‌باشد. علاوه بر این فرض می‌شود که متغیر U_i با استفاده از متغیرهای کمکی توسط فرم خطی زیر ارائه گردد :

$$U_i = -\sum_{k=۱}^{m} \beta_k x_{ki} + \epsilon_i \qquad (۳.۱)$$

که در آن ϵ_i برای $i = ۱, \ldots, n$، متغیرهای تصادفی مستقل و هم توزیع از توزیع تجمعی F می‌باشند.

از مفروضات بالا نتیجه می‌شود که احتمالات تجمعی متغیر مشاهده شدهٔ Y_i برای $i = ۱, \ldots, n$، توسط مدل زیر قابل بیان است:

$$P(Y_i \leq j|\underline{x_i}) = F(\theta_j + \sum_{k=۱}^{m} \beta_k x_{ki}), \quad j = ۱, \ldots, J-۱. \qquad (۴.۱)$$

ازآن جایی که سمت چپ معادله (۴.۱)، مجموع احتمالات

$$P(Y_i = ۱|\underline{x_i}) + ... + P(Y_i = j|\underline{x_i})$$

برای هر واحد می‌باشد، این مدل را مدل تجمعی با تابع توزیع F می‌نامند. به طور مشابه این مدل به مدل آستانه‌ای نیز معروف است زیرا مدل توسط نقاط آستانه‌ای $\theta_۱, ..., \theta_{J-۱}$ مشخص می‌گردد. از معادلهٔ تجمعی بالا می‌توان برای به دست آوردن احتمالات نقطه‌ای به صورت زیر بهره گرفت $(j = ۱, \ldots, J)$:

$$\begin{aligned} p_{ij} = P(Y_i = j|\underline{x_i}) &= P(Y_i \leq j|\underline{x_i}) - P(Y_i \leq j-۱|\underline{x_i}) \\ &= F(\theta_j + \sum_{k=۱}^{m} \beta_k x_{ki}) - F(\theta_{j-۱} + \sum_{k=۱}^{m} \beta_k x_{ki}) \end{aligned}$$

با استفاده از روابط بالا، تابع درستنمایی مدل تجمعی با فرض استقلال میان واحدهای مورد مطالعه به فرم زیر خواهد بود:

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\beta}|\underline{y}, \underline{x}) &= \prod_{i=۱}^{n} P(Y_i = y_i|\boldsymbol{\theta}, \boldsymbol{\beta}) \\ &= \prod_{i=۱}^{n} \prod_{j=۱}^{J} p_{ij}^{y_{ij}} = \prod_{i=۱}^{n} [F(\theta_{y_i} + \sum_{k=۱}^{m} \beta_k x_{ki}) - F(\theta_{y_i-۱} + \sum_{k=۱}^{m} \beta_k x_{ki})]. \end{aligned}$$

که در آن y_{ij} متغیر نشانگری است که مقدار یک را اختیار می‌نماید اگر $y_i = j$ و در غیر این صورت صفر خواهد بود.

انتخاب یک توزیع خاص برای F منجر به یک مدل تجمعی خاص می‌گردد. در زیر سه مدل ممکن با انتخاب توزیع‌های متفاوت مورد بررسی قرار می‌گیرند:

(۱) مدل لوژستیک تجمعی یا مدل بخت‌های نسبتی

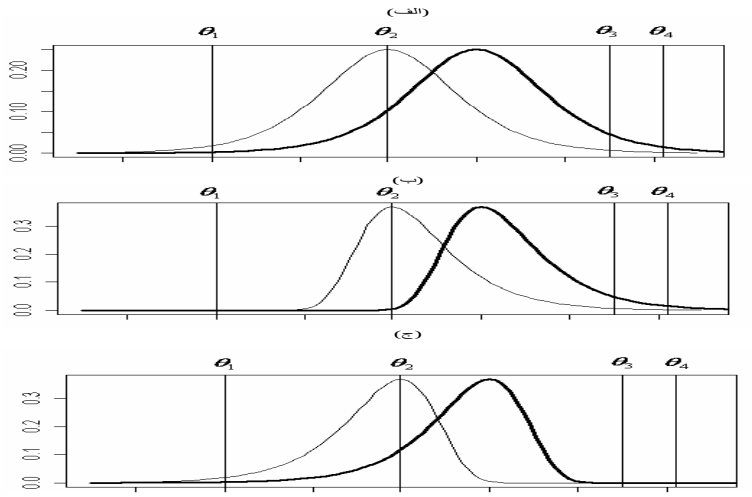
یک انتخاب متداول، توزیع لوژستیک $F(x) = \frac{\exp(x)}{1+\exp(x)},\ x \in \mathcal{R}$ می‌باشد که در نتیجه مدل لوژستیک تجمعی به ازای $j = 1, \ldots, J-1$ به صورت زیر بیان می‌شود:

$$P(Y_i \leq j|\underline{x_i}) = \frac{\exp(\theta_j + \sum_{k=1}^{m} \beta_k x_{ki})}{1 + \exp(\theta_j + \sum_{k=1}^{m} \beta_k x_{ki})} \tag{5.1}$$

رابطهٔ احتمالی فوق را می‌توان به سادگی به فرم معادل زیر نوشت:

$$\log\{\frac{P(Y_i \leq j|\underline{x_i})}{P(Y_i > j|\underline{x_i})}\} = \theta_j + \sum_{k=1}^{m} \beta_k x_{ki}$$

راه دیگر نگاه کردن به مدل (۴.۱) در نظر گرفتن چگالی متغیر پیوسته پنهان U_i یعنی $f = F'$ می‌باشد که در این صورت مکانیسم تولید پاسخ در رابطهٔ (۳.۱) چگالی متغیر U_i را به برش‌هایی تقسیم می‌کند که نقاط برش توسط آستانه‌ها به دست می‌آید و جملهٔ پیشگوی خطی $-\sum_{k=1}^{m} \beta_k x_{ki}$ در رابطهٔ (۳.۱) مقدار جابه‌جایی را در مقیاس متغیر پنهان U_i نشان می‌دهد. بر اساس قوت و جهت این جابه‌جایی، احتمالات رده‌های پاسخ افزایش و یا کاهش می‌یابد. در شکل (۱.۱) برای سه تابع توزیع، چگالی‌ها برای دو زیر جامعه یکی دارای مقدار جابه‌جایی و پیشگوی خطی به فرم $-\sum_{k=1}^{m} \beta_k x_{ki}$ (منحنی پررنگ) و دیگری دارای مقدار جابه‌جایی پیشگوی خطی به فرم $-\sum_{k=1}^{m} \beta_k x^*_{ki}$ (منحنی کمرنگ) رسم شده است. این شکل از طرفی گویای تقارن توزیع لوژستیک، چولگی ازراست توزیع بزرگترین مقدار کرانگین و چولگی از چپ توزیع کوچکترین مقدار کرانگین می‌باشد. از طرف دیگر تغییر احتمالات رده‌ها، میزان جابه‌جایی گفته شده را به وضوح نشان می‌دهد.



شکل ۱.۱:) مقدار جابه‌جایی توزیع‌های مختلف با مقادیر گوناگون پیشگوی خطی، (الف) توزیع لوژستیک ، (ب) توزیع بزرگترین مقدار کرانگین ، (ج) توزیع کوچکترین مقدار کرانگین (منحنی پررنگ: مقدار جابه‌جایی به سمت راست، منحنی کمرنگ: میزان جابه‌جایی به سمت چپ)

مدل لوژستیک تجمعی را مدل بخت‌های نسبتی نیز می‌نامند. این نام به دلیل خاصیتی از مدل (۵.۱) می‌باشد که اگر دو جامعه با متغیرهای کمکی $\underline{x_i}^*$ و $\underline{x_i}$ برای واحد i ام داشته باشیم، نسبت بخت‌های تجمعی این دو جامعه به صورت زیر خواهد بود:

$$\Delta = \frac{P(Y_i \leq j|\underline{x_i})/P(Y_i > j|\underline{x_i})}{P(Y_i \leq j|\underline{x_i}^*)/P(Y_i > j|\underline{x_i}^*)} = \exp\{\sum_{k=1}^{m} \beta_k x_{ki} - \sum_{k=1}^{m} \beta_k x_{ki}^*\} \qquad (6.1)$$

که به ردهٔ پاسخ (j) وابسته نیست. یعنی نسبت بخت‌های تجمعی در هر رده برای دو جامعه یکسان است.

طبق تابع درستنمایی گفته شده برای مدل‌های تجمعی با قرار دادن توزیع لوژستیک به جای F تابع درستنمایی مدل لوژستیک تجمعی به صورت زیر به دست می‌آید:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\underline{y}, \underline{x}) \quad = \quad \prod_{i=1}^{n} \prod_{j=1}^{J} p_{ij}^{y_{ij}}$$

$$= \prod_{i=1}^{n}[\frac{\exp(\theta_{y_i}+\sum_{k=1}^{m}\beta_k x_{ki})}{1+\exp(\theta_{y_i}+\sum_{k=1}^{m}\beta_k x_{ki})}-\frac{\exp(\theta_{y_i-1}+\sum_{k=1}^{m}\beta_k x_{ki})}{1+\exp(\theta_{y_i-1}+\sum_{k=1}^{m}\beta_k x_{ki})}].$$

(۲) مدل پروبیت

اگر تابع توزیع تجمعی نرمال را در رابطهٔ (۴.۱) به کار گیریم، مدل پروبیت به صورت

$$P(Y_i \leq j|\underline{x_i}) = \Phi(\theta_j + \sum_{k=1}^{m}\beta_k x_{ki})$$

برای $j = 1, \ldots, J$ قابل بیان است. ه دلیل شباهت زیاد توزیع لوژستیک و توزیع نرمال (به جز در دم‌ها که توزیع لوژستیک دم‌های کلفت‌تری دارد) غالباً نتایج برآورد این مدل با مدل لوژستیک شباهت فراوانی دارد. تابع درستنمایی مدل نیز به فرم زیر است:

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\beta}|\underline{y},\underline{x}) &= \prod_{i=1}^{n}\prod_{j=1}^{J} p_{ij}^{y_{ij}} \\ &= \prod_{i=1}^{n}[\Phi(\theta_{y_i}+\sum_{k=1}^{m}\beta_k x_{ki})\} - \Phi(\theta_{y_i-1}+\sum_{k=1}^{m}\beta_k x_{ki})]. \end{aligned}$$

(۳) مدل کاکس گروه بندی شده یا نرخ‌های شکست نسبتی

گزینهٔ دیگر برای تابع توزیع F، توزیع کوچکترین مقدار کرانگین $F(x) = 1 - \exp[-\exp(x)], \quad x \in \mathcal{R}$ می‌باشد. با جایگذاری این توزیع در رابطه (۴.۱) مدل زیر برای $j = 1, \ldots, J-1$ و $i = 1, \ldots, n$ به دست می‌آید :

$$P(Y_i \leq j|\underline{x_i}) = 1 - \exp\{-\exp(\theta_j + \sum_{k=1}^{m}\beta_k x_{ki})\} \qquad (۷.۱)$$