Statistical Research Center
J. Stat. Res. Iran 1 (2004) 51-67
مجله‌ی پژوهش‌های آماری ایران
دوره‌ی ۱، شماره‌ی ۱، بهار و تابستان ۱۳۸۳، صص ۵۱-۶۷

مدل‌بندی پاسخ‌های جفت‌شده‌ی ترتیبی

انوشیروان کاظم‌نژاد و فرید زایری

دانشگاه تربیت مدرس

چکیده. حدود ربع قرن از انتشار مقاله‌ی مکولا [۱۱] در مورد مدل‌بندی پاسخ‌های ترتیبی یک‌متغیره می‌گذرد. پس از انتشار این مقاله، مدل پیشنهادی وی به‌تدریج تعمیم یافت، به‌طوری که ما هم‌اکنون قادریم پاسخ‌های چندمتغیره، همبسته و ترتیبی را به‌کمک مدل‌های نسبتاً پیچیده اما کارا تحلیل نموده، ارتباط بین این‌گونه پاسخ‌ها و متغیرهای کمکی مختلف را مورد بررسی قرار دهیم. در این مقاله قصد داریم پیشرفت‌های به وجود آمده در زمینه‌ی مدل‌بندی پاسخ‌های ترتیبی همبسته را با تأکید بر پاسخ‌های دومتغیره‌ی ترتیبی حاصل از مطالعه‌ی اندام‌های جفتی بدن نظیر چشم، گوش، کلیه، دست، پا، و ... بررسی کرده، در پایان، مدلی جدید برای تحلیل داده‌های ترتیبی همبسته معرفی نماییم. هنگامی که داده‌های پاسخ مورد بررسی دارای توزیعی دومتغیره و نامتقارن فرض می‌شوند، این مدل را می‌توان جایگزینی مناسب برای مدل پروبیت تجمعی دومتغیره محسوب نمود. در پایان، به‌عنوان مثالی کاربردی، داده‌های مربوط به وضعیت پریودنتال دانش‌آموزان دبیرستانی شهر تهران به‌کمک این مدل، تحلیل و نتایج حاصل با مدل‌های مشابه مقایسه می‌شود.
© ۱۳۸۳ پژوهشکده‌ی آمار. همه‌ی حقوق محفوظ است.

واژگان کلیدی. پاسخ‌های ترتیبی همبسته؛ توزیع نهانی دومتغیره؛ معادلات براوردگر تعمیم‌یافته؛ مدل‌های خطی تعمیم‌یافته.

۱ مقدمه

مدل‌بندی داده‌های پاسخ ترتیبی، معمولاً پیچیده‌تر از پاسخ‌های دوحالتی و پیوسته به نظر می‌رسد. به همین دلیل، معمول است که آمارشناسان در ابتدا مدل‌هایی برای تحلیل داده‌های دوحالتی یا پیوسته معرفی نموده، سپس این مدل‌ها را برای تحلیل داده‌های ترتیبی، تغییر یا تعمیم می‌دهند. از معروف‌ترین موارد در این زمینه

می‌توان به مدل لوژستیک معمولی اشاره کرد که در ابتدا فقط برای مدل‌بندی پاسخ‌های دوحالتی یک‌متغیره به‌کار برده می‌شد، اما هم‌اکنون تعمیم این مدل را می‌توان برای تحلیل داده‌های چندمتغیره، همبسته و ترتیبی مورد استفاده قرار داد.

با مروری بر مقاله‌های کاربردیِ منتشر شده در سالیان دور، به‌سادگی می‌توان مشاهده کرد که محققان برای تحلیل داده‌های پاسخ ترتیبی و بررسی ارتباط آن‌ها با متغیر مستقل، بیش‌تر از آزمون‌های آماری استفاده می‌کرده‌اند تا مدل‌سازی آماری. در برخی موارد نیز این پاسخ‌های ترتیبی به پاسخ‌های دیگری (مثلاً دوحالتی) تغییر می‌یافتند و سپس با کمک مدل‌های موجود، مورد بررسی قرار می‌گرفتند. در سال ۱۹۸۰ میلادی، مکولا [۱۱] با انتشار مقاله‌ای در مورد مدل‌بندی پاسخ‌های ترتیبی، تحولی در این زمینه ایجاد کرد. با گذشت حدود ربع قرن از انتشار این مقاله، نام آن را همچنان می‌توان در فهرست مرجع‌های اکثر کتاب‌ها و مقاله‌های منتشر شده در مورد تحلیل داده‌های ترتیبی یک‌متغیره یا چندمتغیره دید.

با وجود کارایی مناسب مدل مکولا در تحلیل داده‌های ترتیبی یک‌متغیره، مسئله‌ی مدل‌بندی داده‌های چندمتغیره‌ی ترتیبی و حتا دوحالتی همبسته، همچنان مسئله‌ای پیچیده به نظر می‌رسید. چهار سال بعد، یعنی در سال ۱۹۸۴، رازنر [۱۲] دو مدل مناسب برای تحلیل پاسخ‌های دومتغیره با توزیع دوجمله‌ای و نرمال ارائه کرد. این دو مدل رگرسیون چندگانه، ابزاری مناسب برای تحلیل پاسخ‌های به دست آمده از اندام‌های جفتی بدن در اختیار تحلیلگران قرار می‌دادند، که به‌کمک آن‌ها همبستگی موجود بین مشاهدات پاسخ در مدل‌ها لحاظ می‌شد.

روش معادلات براوردگر لیانگ و زیگر [۸] را می‌توان مهم‌ترین تحول در زمینه‌ی تحلیل داده‌های چندمتغیره‌ی همبسته در سالیان اخیر نامید. در این روش، که بعدها به‌نام روش GEE مشهور شد، نیازی به دانستن توزیع مشترک داده‌های پاسخ چندمتغیره نیست و کافی است که توزیع حاشیه‌ای هر یک از پاسخ‌ها عضوی از خانواده‌ی نمایی فرض شود. این روش به‌طور مستقیم برای تحلیل پاسخ‌های ترتیبی همبسته قابل استفاده نبود، اما لیپ‌شیتس و دیگران [۹] در سال ۱۹۹۴ آن را برای مدل‌بندی داده‌های رسته‌ای همبسته تعمیم دادند. برای براورد پارامترهای رگرسیونی در هر دو روشِ یاد شده، از روش شبه‌درست‌نمایی به‌جای روش حد اکثر درست‌نمایی استفاده می‌شود.

با وجود پیچیده بودنِ تابع درست‌نمایی برای پاسخ‌های چندمتغیره‌ی کیفی همبسته، عده‌ای از آمارشناسان، همچنان بر استفاده از این روش تأکید ورزیده، مدل‌هایی مبتنی بر براوردهای حدا کثر درست‌نمایی را به کار می‌گرفتند. لیپ‌شیتس و دیگران [۱۰] در سال ۱۹۹۰ روشی مبتنی بر حد اکثر درست‌نمایی را برای تحلیل داده‌های جفت‌شده‌ی دوحالتی ارائه کردند. این روش برای تحلیل داده‌های طولی با دو موقعیت اندازه‌گیری یا پاسخ‌های به دست آمده از مطالعات مقطعی کاربرد دارد. با گذشت زمان، مدل‌های مبتنی بر روش حد اکثر درست‌نمایی برای براورد پارامترها، در تحلیل پاسخ‌های دومتغیره‌ی ترتیبی نیز وارد شدند.

© ۱۳۸۳ پژوهشکده‌ی آمار

نخستین مقاله در این زمینه در سال ۱۹۹۵ از سوی کیم [۷] ارائه گردید. وی با ابتکاری جالب توجه، یک توزیع دومتغیره‌ی نرمال را توزیع پایه و نهانی برای داده‌های پاسخ دومتغیره فرض کرد و از این توزیع در تشکیل معادلات درست‌نمایی برای براورد پارامترهای مدل استفاده نمود. مدل کیم به‌نام مدل پروبیت تجمعی دومتغیره مشهور است.

تحلیل داده‌های چندمتغیره‌ی ترتیبی به این نقطه ختم نشد. یک سال بعد، ویلیامسون و دیگران [۱۴] مدلی دومتغیره برای تحلیل پاسخ‌های ترتیبی همبسته ارائه کردند که در آن از نسبت بخت‌های تعمیم‌یافته به‌عنوان معیار همبستگی بین مشاهدات پاسخ استفاده می‌شد. چند سال بعد، مدل‌های مبتنی بر اثرهای تصادفی [۴] نیز وارد عرصه‌ی تحلیل پاسخ‌های ترتیبی همبسته شدند. طرفداران تحلیل‌های بیزی نیز سرانجام به میدان آمده، مدلی مبتنی بر نمونه‌گیری MCMC و الگوریتم Gibbs برای تحلیل داده‌های ترتیبی دومتغیره ارائه کردند. این مقاله در سال ۲۰۰۲ از سوی بیسواس و داس [۳] منتشر شد. با گسترش روش‌های مختلف برای تحلیل داده‌های ترتیبی چندمتغیره، نیازی مبرم به ایجاد امکانات نرم‌افزاری برای برازش این مدل‌ها احساس شد. ویلیامسون و دیگران [۱۵] در سال ۱۹۹۹ طی مقاله‌ای دو نرم‌افزار رایانه‌ای برای تحلیل داده‌های پاسخ رسته‌ایِ همبسته معرفی کردند. پس از سال ۲۰۰۰، نرم‌افزار قدرتمند SAS، امکانات لازم را برای برازش گروه عمده‌ای از مدل‌های مناسب برای تحلیل داده‌های چندمتغیره‌ی همبسته شامل براوردهای GEE و حد اکثر درست‌نمایی فراهم آورد. با این حال، برازش برخی از انواع این مدل‌ها، همچنان نیازمند طراحی برنامه‌های پیچیده‌ی رایانه‌ای در محیط برنامه‌نویسی نرم‌افزارهایی نظیر SAS، S-Plus، و . . . است.

در این مقاله امکان بحث در مورد همه‌ی روش‌های یاد شده وجود ندارد. از آن‌جایی که مدل ارائه شده از سوی ما، با فرض یک توزیع پیوسته‌ی نهانی برای داده‌های جفت‌شده‌ی ترتیبی بنا می‌شود، در بخش بعد فقط به توصیف مدل‌های مبتنی بر توزیع‌های دومتغیره‌ی نهانی اکتفا می‌کنیم. سپس مهم‌ترین مدل دومتغیره‌ی ارائه شده در سال‌های اخیر (با فرض یک توزیع دومتغیره‌ی نهانی) را مورد بحث قرار می‌دهیم. همان‌طور که پیش از این گفته شد، این مدل را کیم ارائه کرده است و در آن از توزیع نرمال دومتغیره‌ی استاندارد به‌عنوان متغیر نهانی استفاده می‌شود. با وجود مناسب بودن این مدل برای تحلیل طیف وسیعی از داده‌های جفت‌شده‌ی ترتیبی، هنگامی که پاسخ دومتغیره حالت متقارنی نداشته باشد، به‌کارگیری این مدل (به‌علت فرم متقارن توزیع نرمال دومتغیره) چندان منطقی به نظر نمی‌آید. برای رفع این نقیصه، در بخش ۳٫۲ مدلی ارائه می‌کنیم که در آن، توزیع دومتغیره‌ی نهانی از فرم نامتقارنی برخوردار بوده، نتیجه انتخابی مناسب برای تحلیل برخی از پاسخ‌های دومتغیره‌ی نامتقارن است.

پس از توصیف این مدل، روش‌های مورد نیاز برای براورد پارامترهای آن بیان شده، سپس داده‌های حاصل از یک مطالعه‌ی اپیدمیولوژیک مربوط به وضعیت پریودنتال دانش‌آموزان دبیرستانی شهر تهران را

© ۱۳۸۳ پژوهشکده‌ی آمار

به‌وسیله‌ی این مدل، تحلیل می‌کنیم. در پایان، نتایج حاصل از برازش مدل پیشنهادی، با نتایج حاصل از مدل پروبیت تجمعی دومتغیره‌ی کیم و همچنین مدل مبتنی بر براورد GEE مقایسه خواهد شد.

۲ مدل‌های رگرسیونی برای پاسخ‌های ترتیبی یک‌متغیره

در این بخش به‌طور مختصر، مدل‌های رگرسیونی با پاسخ‌های ترتیبی یک‌متغیره را مورد بحث قرار می‌دهیم. در همه‌ی این مدل‌ها فرض بر این است که یک متغیر نهانی پیوسته وجود دارد که به‌طور مستقیم قابل مشاهده نیست، اما توزیع آن به‌عنوان توزیع پایه‌ی متغیر پاسخ در نظر گرفته می‌شود.

فرض کنیم y^* یک متغیر تصادفی پیوسته است که به‌طور مستقیم قادر به مشاهده‌ی آن نیستیم. در عوض، فقط مشاهده می‌کنیم که آیا y^* به فواصل مجاوری نظیر (θ_{h-1}, θ_h)، $h = 1, \ldots, H$، با فرض $\theta_0 = -\infty$ و $\theta_H = +\infty$، تعلق دارد یا خیر. نقاط برش، یعنی $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{h-1})$، معمولاً نامعلوم هستند و در عوض، متغیر تصادفی ترتیبی y به‌صورت زیر قابل مشاهده است:

$$\theta_{h-1} < y^* \leqslant \theta_h \qquad \Rightarrow \qquad y = h.$$

در حالت کلّی فرض کنیم N فرد در مطالعه حضور دارند و y_i^*، $(i = 1, \ldots, N)$، نشان‌دهنده‌ی متغیر پیوسته‌ی نهانی پاسخ برای iامین فرد تحت مطالعه با یک رگرسیون خطی روی متغیرهای کمکی $\boldsymbol{x}_i$ است. به‌طور دقیق‌تر، فرض بر این است که $\frac{y_i^* - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}$ ، متغیرهای تصادفی با تابع توزیع مشترک F هستند. در این‌جا σ پارامتر مقیاس و $\boldsymbol{\beta}$ برداری از ضرایب رگرسیونی فرض می‌شود. بدون از دست رفتن کلّیت مطلب، فرض کنیم $\sigma = 1$. در این صورت، متغیر پاسخ رسته‌ای y_i برای iامین فرد، دارای تابع چگالی احتمال زیر خواهد بود:

$$\Pr(y_i = h) = \pi_h(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{x}_i),$$

که در آن:

$$\begin{aligned}\pi_h(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{x}_i) &= \Pr(\theta_{h-1} < y_i^* \leqslant \theta_h) \\ &= F(\theta_h - \boldsymbol{x}_i'\boldsymbol{\beta}) - F(\theta_{h-1} - \boldsymbol{x}_i'\boldsymbol{\beta}).\end{aligned}$$

حال، بر حسب احتمالات تجمعی $\gamma_h = \sum_{s=1}^{h} \pi_s$، می‌توان مدل رگرسیونی را به‌صورت زیر نوشت:

$$F^{-1}\{\gamma_h(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{x}_i)\} = \theta_h - \boldsymbol{x}_i'\boldsymbol{\beta}.$$

در این‌جا می‌توان یک توزیع نهانی پیوسته‌ی F به‌عنوان تابع ربط برای احتمال‌های تجمعی فوق فرض کرد. با این کار، در واقع، مدل به‌صورت یک مدل خطی تعمیم‌یافته نوشته خواهد شد. معمول‌ترین انتخاب در

© ۱۳۸۳ پژوهشکده‌ی آمار

این زمینه، تابع توزیع لوژستیک است که مدل لوژستیک را نتیجه می‌دهد:

$$\log\left(\frac{\gamma_h}{1-\gamma_h}\right) = \theta_h - \boldsymbol{x}_i'\boldsymbol{\beta}.$$

تابع توزیع گامبل، انتخاب دیگری است که مدل مکمل log-log را به دست می‌دهد:

$$\log\{-\log(1-\gamma_h)\} = \theta_h - \boldsymbol{x}_i'\boldsymbol{\beta}.$$

همچنین انتخاب تابع توزیع نرمال، مدل پروبیت را نتیجه خواهد داد:

$$\varphi^{-1}(\gamma_h) = \theta_h - \boldsymbol{x}_i'\boldsymbol{\beta},$$

که در آن $\varphi^{-1}(\cdot)$ وارون تابع توزیع نرمال استاندارد است.

به‌علت وجود تقارن در توزیع‌های لوژستیک و نرمال، معمولاً مدل‌های لوژستیک و پروبیت، نتایجی مشابه یکدیگر خواهند داشت. اما از مدل‌هایی با تابع پیوند log-log در عمل، هنگامی استفاده می‌شود که داده‌های پاسخ، حالت متقارن نداشته باشند [۲].

پیش از ارائه‌ی روش‌های مناسب برای تحلیل چندمتغیره‌ی پاسخ‌های جفت‌شده‌ی ترتیبی، تحلیلگران در ابتدا پاسخ‌های چندمتغیره را با روش‌هایی چون گرفتن میانگین یا میانه، در نظر گرفتن پاسخ ماکسیمم برای هر فرد، یا روش‌های دیگری مانند روش تأکید بر اندام بیمارتر [۵] به پاسخی یک‌متغیره تبدیل کرده، سپس با استفاده از مدل‌های توصیف‌شده در این بخش، داده‌ها را تحلیل می‌کردند. اما به‌تدریج با مطرح شدن روش‌هایی برای تحلیل پاسخ‌های دومتغیره یا چندمتغیره، مدل‌های یک‌متغیره اهمیت خود را در تحلیل این‌گونه پاسخ‌ها از دست دادند. آنانث و کلاین‌باوم [۱] مقاله‌ی مروری مبسوطی در مورد سیر تحولات روش‌های مربوط به تحلیل داده‌های ترتیبی یک‌متغیره ارائه کرده‌اند. اشکال مهمِ به‌کارگیری مدل‌های یک‌متغیره برای تحلیل پاسخ‌های چندمتغیره‌ی همبسته این است که با تبدیل یک پاسخ چندمتغیره به مشاهده‌ای یک‌متغیره، عملاً امکان احتساب همبستگی بین مشاهدات پاسخ در مدل را از دست می‌دهیم. همچنین واضح است که تبدیل پاسخ‌های چندمتغیره به یک‌متغیره ممکن است موجب بروز تغییرات کلّی در ساختار روابط پاسخ مورد بررسی و متغیرهای تبیینی گردد. در بخش‌های بعدی، مدل‌هایی را معرفی می‌کنیم که در آن‌ها امکان تحلیل پاسخ‌های دومتغیره‌ی همبسته و ترتیبی با احتساب همبستگی بین مشاهدات پاسخ وجود دارد.

© ۱۳۸۳ پژوهشکده‌ی آمار

۳ مدل‌های رگرسیونی برای پاسخ‌های جفت‌شده‌ی ترتیبی

۳٫۱ مدل پروبیت تجمعی برای پاسخ‌های دومتغیره‌ی متقارن

همان‌طور که در فصل پیش اشاره شد، مدل پروبیت تجمعی دومتغیره در سال ۱۹۹۵ از سوی کیم ارائه گردید. قبل از توصیف این مدل، بهتر است اشاره‌ای به ساختار داده‌های جفت‌شده در حالت کلّی داشته باشیم.

در مطالعه‌ای پزشکی مربوط به اندام‌های جفتی بدن، فرض کنیم y_{i1} و y_{i2} نشان‌دهنده‌ی پاسخ‌های دومتغیره‌ی ترتیبی، به‌ترتیب برای اندام راست و چپ (یا در مطالعات دندان‌پزشکی، بالا و پایین) فرد iام باشد. بنا بر این، پاسخ دومتغیره‌ی فرد iام را می‌توان به‌صورت $\boldsymbol{y}_i = (y_{i1}, y_{i2})'$ نمایش داد. همچنین فرض کنیم y_{i1} و y_{i2} مقادیر ۱ و ۲ و ... و H را اختیار کنند. حال همانند حالت یک‌متغیره، متغیرهای پیوسته و غیر قابل مشاهده‌ی y^*_{i1} و y^*_{i2} را در نظر می‌گیریم. بار دیگر، نقاط برش θ_0، θ_1، ...، θ_H، ($\theta_0 = -\infty$ و $\theta_H = +\infty$) وجود دارند، به‌طوری که

$$\theta_{h-1} < y^*_{ij} \leqslant \theta_h \qquad \Rightarrow \qquad y_{ij} = h.$$

توجه کنید که در این‌جا نقاط برش یکسانی برای هر دو حاشیه در نظر گرفته شده است. فرض یکسان بودن حاشیه‌ها برای اندام‌های جفتی بدن، فرضی غیرمنطقی به نظر نمی‌رسد. با این حال، چنانچه ماهیت داده‌ها ایجاب کند، می‌توان نقاط برش θ_h، ($h = 1, \ldots, H-1$) را با θ_{jh}، ($j = 1, 2$ و $h = 1, \ldots, H-1$) جایگزین نمود.

در مطالعات مربوط به اندام‌های جفتی بدن، دو نوع متغیر کمکی را می‌توان در نظر گرفت: متغیرهای کمکی **مختص فرد** و متغیرهای کمکی **مختص اندام**. اگر $\boldsymbol{x}_i$ نشان‌دهنده‌ی برداری p-بعدی از متغیرهای کمکی مختص فرد iام، و $\boldsymbol{x}_{i1}$ و $\boldsymbol{x}_{i2}$ نشان‌دهنده‌ی بردارهای q-بعدی مختص اندام‌های راست و چپ مربوط به همین فرد باشند، آن‌گاه ماتریس متغیرهای کمکی (ماتریس طرح) مربوط به iامین فرد را می‌توان به‌صورت زیر نمایش داد:

$$\boldsymbol{X}'_i = \begin{pmatrix} \boldsymbol{x}'_i & \boldsymbol{x}'_{i1} \\ \boldsymbol{x}'_i & \boldsymbol{x}'_{i2} \end{pmatrix}.$$

با این تعریف، این بار فرض کنیم $\frac{y^*_i - \boldsymbol{X}'_i\boldsymbol{\beta}}{\sigma}$ متغیرهای تصادفی مستقل با توزیع یکسان و تابع توزیع دومتغیره‌ی مشابه F_ρ باشند، به‌طوری که پارامتر ρ همبستگی بین y^*_{i1} و y^*_{i2} را مشخص می‌کند. همچنین، $\boldsymbol{y}^*_i = (y^*_{i1}, y^*_{i2})'$ نشان‌دهنده‌ی بردار متغیرهای نهانی مختص فرد iام است. در این‌جا، $\boldsymbol{\beta} = (\boldsymbol{\delta}', \varepsilon')'$

© ۱۳۸۳ پژوهشکده‌ی آمار

بردار پارامترهای رگرسیونی مختص فرد و مختص اندام را نمایش می‌دهد و σ همان پارامتر مقیاس معمولی است. بار دیگر فرض کنیم $\sigma = ۱$.

از آن‌جایی که متغیرهای پاسخ مربوط به اندام‌های راست و چپ، هر یک ممکن است در یکی از طبقات ۱ تا H قرارگیرند، پاسخ دومتغیره‌ی هر فرد ممکن است در یکی از خانه‌های یک جدول $H \times H$ قرارگیرد و در نتیجه پاسخ کل این N فرد را می‌توان در این جدول $H \times H$ خلاصه کرد.

حال فرض می‌کنیم $\pi_{hg}(\boldsymbol{\omega};\ \boldsymbol{X}_i)$ نشان‌دهنده‌ی احتمال وقوع (h,g)امین خانه‌ی جدول توافقی $H \times H$ یاد شده باشد. به عبارت دیگر،

$$\pi_{hg}(\boldsymbol{\omega};\boldsymbol{X}_i) = \Pr(y_{i۱} = h,\ y_{i۲} = g)$$

برای هر $h, g = ۱, \ldots, H$. در این‌جا $\boldsymbol{\omega} = (\boldsymbol{\theta}', \rho, \boldsymbol{\delta}', \varepsilon')'$ بردار $(H + p + q)$-بعدی مربوط به پارامترهای مدل است. در واقع، در فرایند مدل‌سازی باید $H - ۱$ نقطه‌ی برش، یک پارامتر همبستگی، p پارامتر مربوط به متغیرهای کمکی مختص فرد، و q پارامتر مربوط به متغیرهای کمکی مختص اندام براورد شود.

بار دیگر به‌دلیل ماهیت ترتیبی متغیر پاسخ دومتغیره، بهتر است احتمال‌های تجمعی γ_{hg}، $\gamma_{hg} = \sum_{m=۱}^{h}\sum_{n=۱}^{g}\pi_{mn}$، را در مدل وارد کنیم. هنگامی که فرض تقارن توزیع پاسخ‌های حاشیه‌ایِ به دست آمده از هر اندام، فرضی معقول باشد، معمول‌ترین انتخاب ممکن برای متغیر پایه‌ی نهانی، توزیع نرمال استاندارد دومتغیره است [۷]. به عبارت بهتر، می‌توان نوشت:

$$\begin{aligned}\gamma_{hg}(\boldsymbol{\omega};X_i) &= F_\rho(\theta_h - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۱}'\varepsilon,\ \theta_g - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۲}'\varepsilon)\\ &= \varphi_\rho(\theta_h - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۱}'\varepsilon,\ \theta_g - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۲}'\varepsilon),\end{aligned}$$

که در آن φ_ρ تابع توزیع نرمال دومتغیره‌ی استاندارد است. بنا بر این، احتمال (h,g)امین خانه‌ی جدول $H \times H$ یاد شده را می‌توان به‌صورت زیر نشان داد:

$$\begin{aligned}\pi_{hg}(\boldsymbol{\omega};\boldsymbol{X}_i) &= \varphi_\rho(\theta_h - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۱}'\varepsilon,\ \theta_g - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۲}'\varepsilon)\\ &\quad - \varphi_\rho(\theta_h - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۱}'\varepsilon,\ \theta_{g-۱} - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۲}'\varepsilon)\\ &\quad - \varphi_\rho(\theta_{h-۱} - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۱}'\varepsilon,\ \theta_g - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۲}'\varepsilon)\\ &\quad + \varphi_\rho(\theta_{h-۱} - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۱}'\varepsilon,\ \theta_{g-۱} - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i۲}'\varepsilon).\end{aligned}$$

به‌منظور براورد پارامترهای مدل بالا، تابع لگاریتم درست‌نمایی برای داده‌های دومتغیره‌ی ترتیبی به‌صورت

© ۱۳۸۳ پژوهشکده‌ی آمار

زیر خواهد بود:

$$\ell(\boldsymbol{\omega}) = \sum_{i=1}^{N} \sum_{h=1}^{H} \sum_{g=1}^{H} I_{i1h} I_{i2g} \log \pi_{hg}(\boldsymbol{\omega}; \boldsymbol{X}_i),$$

که در آن، $I_{ijh} = 1$ اگر پاسخ مربوط به jامین اندام فرد iام در رسته‌ی h قرار داشته باشد، و در غیر این صورت $I_{ijh} = 0$. براورد کردن پارامترها در این مدل، نیازمند روشی تکراری نظیر نیوتون-رافسون برای حل معادله‌ی $\frac{\partial \ell}{\partial \boldsymbol{\omega}} = 0$ و ارزیابی ماتریس مشتق‌های جزئی مرتبه‌ی دوم (Hessian matrix) به‌صورت $-\frac{\partial^2 \ell}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}'}$ برای به دست آوردن خطاهای معیار پارامترهای براورد شده است.

۳٫۲ مدل تجمعی دومتغیره برای پاسخ‌های نامتقارن

فرض متقارن بودن توزیع‌های حاشیه‌ای متغیرهای پاسخ ترتیبی، فرضی است که در عمل، همواره برقرار نخواهد بود. بنا بر این، به‌کارگیری توزیع نرمال دومتغیره به‌عنوان متغیر پایه‌ی نهانی برای انواع داده‌های پاسخِ به دست آمده از اندام‌های جفتی بدن، ممکن است در برخی موارد منجر به استنباط‌هایی گمراه‌کننده در باره‌ی متغیر پاسخِ مورد بررسی شود.

برای رفع نقیصه‌ی یاد شده، ما مدل دیگری مبتنی بر یک توزیع دومتغیره‌ی نامتقارن نهانی پیشنهاد می‌کنیم. این توزیع در سال ۱۹۷۸ از سوی ساترت‌وِیت و هاچینسون [۱۳] پیشنهاد شده است. در این بخش، نخست این توزیع دومتغیره را به‌طور مختصر مورد بحث قرار داده، سپس مدل پیشنهادی خود را ارائه می‌کنیم.

در سال ۱۹۶۱، گامبل [۶] دو حالت از توزیع لوژستیک دومتغیره ارائه کرد، که یکی از آن‌ها دارای تابع توزیع زیر بود:

$$\psi(x, y) = \left(1 + e^{-x} + e^{-y}\right)^{-1}.$$

اشکال بزرگ این توزیع دومتغیره، فقدان پارامترهای مرتبط با همبستگی بین دو متغیر تصادفی X و Y است. برای رفع این اشکال، ساترت‌وِیت و هاچینسون توزیع فوق را به‌صورت زیر اصلاح کردند:

$$\psi(x, y) = \left(1 + e^{-x} + e^{-y}\right)^{-v},$$

که در آن، $v > 0$ پارامتری است که به‌کمک آن می‌توان ضریب همبستگی بین X و Y را محاسبه کرد. این دو آمارشناس در مقاله‌ی خود، خواص مهم این توزیع (نظیر توزیع‌های حاشیه‌ای، امید ریاضی و واریانس، کوواریانس بین X و Y، و ...) را ارائه نمودند. بر این اساس، آنان ضریب همبستگی بین X و Y را

© ۱۳۸۳ پژوهشکده‌ی آمار

به‌صورت زیر به دست آوردند:

$$\rho = \frac{\zeta(2, v)}{\zeta(2, v) + \frac{1}{6}\pi^2},$$

که در آن، $\zeta(s, a) = \sum_{m=0}^{\infty}(m + a)^{-s}$ تابع زتای ریمان (zeta Riemann function) است. بدیهی است در این توزیع، همواره داریم $0 \leqslant \rho \leqslant 1$. توزیع فوق بر خلاف توزیع نرمال دومتغیره، دارای شکلی نامتقارن است.

با تعاریفی کاملاً مشابه با مدل پروبیت تجمعی دومتغیره برای احتمال‌های π_{hg}، می‌توان مدل زیر را با فرض متغیر نهانیِ توصیف‌شده در این بخش، بر اساس احتمال‌های تجمعی γ_{hg} بنا کرد:

$$\begin{aligned}\gamma_{hg}(\boldsymbol{\omega}; \boldsymbol{X}_i) &= F_\rho(\theta_h - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i1}'\boldsymbol{\varepsilon},\ \theta_g - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i2}'\boldsymbol{\varepsilon}) \\ &= \psi_v(\theta_h - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i1}'\boldsymbol{\varepsilon},\ \theta_g - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{i2}'\boldsymbol{\varepsilon}),\end{aligned}$$

که در آن، ψ_ρ تابع توزیع پیشنهاد شده از سوی ساترت‌وِیت و هاچینسون است. در این‌جا نیز $\boldsymbol{\omega} = (\boldsymbol{\theta}', \rho, \boldsymbol{\delta}', \boldsymbol{\varepsilon}')'$ بردار پارامترهای مدل را نمایش می‌دهد. پس از براورد کردن پارامتر همبستگی v در فرایند مدل‌سازی، به‌سادگی می‌توان مقدار ضریب همبستگی ρ را به‌کمک آن براورد نمود. برای براورد کردن پارامترهای مدل، با فرض

$$z_{sj} = \theta_s - \boldsymbol{x}_i'\boldsymbol{\delta} - \boldsymbol{x}_{ij}'\boldsymbol{\varepsilon}$$

برای $j = 1, 2$، می‌توان مدل را در حالت کلّی به‌صورت زیر نوشت:

$$\begin{aligned}\pi_{hg}(\boldsymbol{\omega}; \boldsymbol{x}_i) &= \psi_v(z_{h1}, z_{g2}) - \psi_v(z_{h1}, z_{(g-1),2}) \\ &\quad - \psi_v(z_{(h-1),1}, z_{g2}) + \psi_v(z_{(h-1),1}, z_{(g-1),2}).\end{aligned}$$

تشکیل تابع لگاریتم درست‌نمایی، بر حسب مقدار کاملاً مشابه مدل پروبیت تجمعی دومتغیره خواهد بود.

برای براورد کردن پارامترهای این مدل، از محیط برنامه‌نویسی نرم‌افزار S-Plus استفاده کردیم. روال nlminb در این نرم‌افزار، ابزاری مناسب برای به دست آوردن براورد پارامترهای مدل‌های مختلف غیر خطی به‌کمک روش حد اکثر درست‌نمایی است. برای به دست آوردن براورد پارامترها، روال nlminb از روش تکراری گاوس-نیوتون استفاده می‌کند. رایانه‌های جدید با قدرت پردازش بالا قادرند عملیات فرایند مدل‌سازی را در طی چند دقیقه انجام دهند.

© ۱۳۸۳ پژوهشکده‌ی آمار

۴ نتایج حاصل از یک تحقیق دندان‌پزشکی

در این بخش، داده‌های حاصل از یک مطالعه‌ی اپیدمیولوژیک در مورد وضعیت پریودنتال دانش‌آموزان دبیرستانی شهر تهران به‌کمک روش‌های توصیف‌شده، مورد تحلیل قرار می‌گیرد. نمونه‌ای تصادفی به اندازه‌ی ۸۶۷ دانش‌آموز ۱۵-۱۹ ساله‌ی تهرانی انتخاب شده و وضعیت لثه‌ی آنان به‌وسیله‌ی شاخصی بهداشتی موسوم به شاخص CPI (Community Periodontal Index) مورد بررسی قرارگرفته است. CPI شاخصی ترتیبی با مقادیر زیر است: ۰= لثه‌ی سالم، ۱= خون‌ریزی در لثه، ۲= وجود جرم در لثه، ۳= وجود پاکت با عمق ۴-۶ میلی‌متر، ۴= وجود پاکت با عمق بیش از ۶ میلی‌متر.

مقدار CPI معمولاً در شش دندان شاخص، که نشان‌دهنده‌ی شش بخش فک بالا و پایین دهان هر فرد است، اندازه‌گیری می‌شود. این شش بخش (sextant) عبارت‌اند از: سمت راست فک بالا، قسمت میانی فک بالا، سمت چپ فک بالا، سمت راست فک پایین، قسمت میانی فک پایین، و سمت چپ فک پایین. بنا بر این، برای هر فرد، شش مقدار ترتیبی همبسته به‌عنوان پاسخ در اختیار داریم. از آن‌جایی که هدف ما مدل‌بندی داده‌های ترتیبی دومتغیره است، در ابتدا ضریب همبستگی اسپیرمن برای سمت راست بالا و پایین، میانه‌ی بالا و پایین، و سمت چپ بالا و پایین محاسبه شد. چون بالاترین مقدار ضریب همبستگی اسپیرمن در داده‌های مربوط به بخش میانی فک بالا و پایین به دست آمد ($r = ۰/۷۴۱$ با $P < ۰/۰۰۰۱$)، ما تحلیل خود را بر این دو نقطه متمرکز خواهیم کرد. جدول ۱ پاسخ دومتغیره را به‌طور خلاصه نمایش می‌دهد.

جدول ۱. وضعیت پریودنتال (CPI) بخش میانی فک بالا و پایین در نمونه‌ی دانش‌آموزان تهران

	فک پایین				
	سالم	خون‌ریزی	جرم	پاکت کم‌عمق	پاکت عمیق	جمع
فک بالا						
سالم	۲۹۷	۸۸	۲۱	۴	۰	۴۱۰
خون‌ریزی	۲۷	۱۳۱	۶۵	۹	۰	۲۳۲
جرم	۵	۱۴	۱۳۳	۱۷	۰	۱۶۹
پاکت کم‌عمق	۳	۶	۱۹	۲۷	۱	۵۶
پاکت عمیق	۰	۰	۰	۰	۰	۰
جمع	۳۳۲	۲۳۹	۲۳۸	۵۷	۱	۸۶۷

از بین متغیرهای کمکی مختلف، ما تأثیر شش متغیر کمکی مختص فرد، و فقط یک متغیر کمکی مختص اندام را بر وضعیت پریودنتال دانش‌آموزان مورد بررسی قرار می‌دهیم. متغیرهای کمکی مختص فرد عبارت‌اند از: جنس (۱= زن، ۲= مرد)، تحصیلات پدر (۱= کم‌تر از دیپلم، ۲= دیپلم و بالاتر)، تحصیلات

© ۱۳۸۳ پژوهشکده‌ی آمار

مادر (۱= کم‌تر از دیپلم، ۲= دیپلم و بالاتر)، استفاده از مسواک (۱= هرگز یا به‌طور نامرتب، ۲= به‌طور مرتب، حد اقل روزی یک بار)، استفاده از نخ دندان (۱= هرگز یا به‌طور نامرتب، ۲= به‌طور مرتب، حد اقل روزی یک بار)، ویزیت توسط دندان‌پزشک (۱= فقط در مواقع اضطراری، ۲= به‌طور مرتب به‌منظور پیش‌گیری). همچنین، فقط وجود پلاک دندانی قابل مشاهده در قسمت میانی فک بالا یا فک پایین به‌عنوان متغیر کمکی مختص اندام در نظرگرفته شد. در هر یک از فک‌های بالا و پایین، این متغیر به این صورت ثبت شد: پلاک قابل مشاهده (۱= وجود، ۲= نبود).

پیش از برازش مدل‌های دومتغیره به این داده‌ها، مدل‌های لجیت، پروبیت و log-log برازانده شده به پاسخ‌های حاشیه‌ایِ به دست آمده از فک‌های بالا و پایین را مقایسه کردیم. آماره‌های نیکویی برازش برای مدل با پیوند نامتقارن (log-log به‌خصوص برای فک پایین) به‌طور معناداری بهتر از نتایج برازش مدل‌های لوجیت و پروبیت به دست آمد. در مرحله‌ی بعد، برازش سه مدل چندمتغیره‌ی زیر را مورد بررسی قرار دادیم:

۱. مدل لوجیت تجمعی که پارامترهای آن به‌روش GEE براورد می‌شوند؛

۲. مدل پروبیت تجمعی که پارامترهای آن با روش حد اکثر درست‌نمایی براورد می‌شوند (مدل با تابع پیوند متقارن یا مدل کیم)؛

۳. مدل ارائه شده در بخش ۳٫۲ (مدل با تابع پیوند نامتقارن).

به‌علت کوچک بودن اندازه‌ی نمونه برای رسته‌ی دارای CPI=۴، رسته‌های ۳ و ۴ داده‌های پاسخ را ادغام کردیم. بنا براین، پاسخ‌های حاشیه‌ای به این صورت در مدل وارد شدند: ۰= سالم، ۱= خون‌ریزی، ۲= جرم، ۳= پاکت با عمق بیش از ۴ میلی‌متر. جدول ۲ نتایج برازش این سه مدل را نمایش می‌دهد. همان‌طور که از مقادیرِ منهای دو برابر لگاریتم درست‌نمایی مشخص است، بهترین برازش، همان‌گونه که انتظار می‌رفت، مربوط به مدل نامتقارن و پس از آن متعلق به مدل کیم و در آخر، مربوط به مدل مبتنی بر براوردهای GEE است. بدیهی است که براورد پارامترها، مقادیر خطاهای معیار و پی‌مقدار برای سه مدل، یکسان نخواهد بود. بر طبق نتایجِ به دست آمده از مدل نامتقارن، کلّیه‌ی متغیرهای کمکی دارای تأثیری معنادار بر وضعیت پریودنتال دانش‌آموزان دبیرستانی شهر تهران هستند. مطابق مدل کیم، متغیر تحصیلات پدر، یگانه متغیر کمکی غیر معنادار در مدل است. همچنین بر طبق نتایج حاصل از برازش GEE، دو متغیر تحصیلات پدر و تحصیلات مادر، تأثیر معناداری بر وضعیت پریودنتال دانش‌آموزان تهرانی نداشته‌اند. تفسیر ضرایب رگرسیونی براورد شده را می‌توان بر حسب مقیاس متغیر نهانی انجام داد. به‌عنوان مثال، در مدل نامتقارن، از آن‌جایی که اثر استفاده از مسواک به‌صورت $\hat{\delta}_۴ = ۱٫۷۵۵$ و اثر استفاده از نخ دندان به‌صورت $\hat{\delta}_۵ = ۱٫۵۴۶$ براورد شده است، اگر این دو اثر را به‌طور توأم در نظر بگیریم، یعنی ۱٫۷۵۵+۱٫۵۴۶=۳٫۳۰۱، می‌توان

© ۱۳۸۳ پژوهشکده‌ی آمار

جدول ۲. نتایج برازش مدل‌های رگرسیونی چندمتغیره برای داده‌های پریودنتال

پارامتر	مدل نامتقارن			مدل کیم			مدل GEE		
	براورد	SE	P	براورد	SE	P	براورد	SE	P
θ_1	۱۱٫۰۵۴	۱٫۲۳۵	†	۵٫۱۱۲	۱٫۰۹۱	†	−۲٫۲۲۴	۰٫۱۱۵	†
θ_2	۱۳٫۱۸۴	۱٫۵۴۱	†	۷٫۴۹۹	۰٫۱۳۲	†	−۱٫۲۸۷	۰٫۱۰۳	†
θ_3	۱۵٫۲۶۹	۱٫۵۳۸	†	۹٫۴۲۵	۰٫۳۲۵	†	۰٫۰۹۵	۰٫۱۰۲	†
δ_1 : جنس	۰٫۹۹۹	۰٫۶۳۲	۰٫۰۳۱	۰٫۷۳۳	۰٫۲۷۸	۰٫۰۴۱	۰٫۴۲۰	۰٫۰۷۲	< ۰٫۰۰۱
δ_2 : تحصیلات پدر	۰٫۳۲۹	۰٫۱۲۷	۰٫۰۳۸	۰٫۱۹۸	۰٫۰۸۳	۰٫۰۶۱	۰٫۲۰۶	۰٫۱۶۰	۰٫۱۹۸
δ_3 : تحصیلات مادر	۰٫۹۱	۰٫۲۸۹	< ۰٫۰۰۱	۰٫۵۵۲	۰٫۱۶۰	< ۰٫۰۰۱	۰٫۴۴۶	۰٫۲۳۸	۰٫۰۶۰
δ_4 : مسواک	۱٫۷۵۵	۰٫۲۱۲	< ۰٫۰۰۱	۰٫۸۳۰	۰٫۱۱۲	< ۰٫۰۰۱	۰٫۶۵۳	۰٫۰۷۹	< ۰٫۰۰۱
δ_5 : نخ دندان	۱٫۵۴۶	۰٫۱۹۷	< ۰٫۰۰۱	۰٫۹۵۶	۰٫۱۹۹	< ۰٫۰۰۱	۰٫۷۲۹	۰٫۱۴۱	< ۰٫۰۰۱
δ_6 : ویزیت	۱٫۱۳۹	۰٫۲۷۴	< ۰٫۰۰۱	۰٫۷۷۳	۰٫۱۹۴	< ۰٫۰۰۱	۰٫۵۱۳	۰٫۰۷۹	< ۰٫۰۰۱
ε_1 : پلاک	۱٫۵۸۱	۰٫۱۱۳	< ۰٫۰۰۱	۱٫۰۰۵	۰٫۰۹۱	< ۰٫۰۰۱	۱٫۱۷۵	۰٫۰۶۴	< ۰٫۰۰۱
پارامتر همبستگی	۰٫۴۴۷	۰٫۰۹۸	†	۰٫۸۰۳	۰٫۱۳۱	†	†	†	†
-2log(likelihood)	۳۲۵۶٫۲۸۳			۳۳۰۱٫۱۶۸			۳۵۲۲٫۲۲۷		

† محاسبه نمی‌شود.

چنین استنباط کرد که تغییر از استفاده به عدم استفاده از مسواک و نخ دندان به‌طور توأم، باعث تغییری به‌اندازه‌ی ۳٫۳۰۱ در مقیاس متغیر نهانی می‌شود. با توجه به مقادیر عرض از مبدأ براورد شده، این میزان تغییر می‌تواند وضعیت لثه‌ی یک دانش‌آموز را از حالت خون‌ریزی ($\hat{\theta}_1 \leqslant y^*_{ij} = ۱۳٫۱۸۴$) به حالت وجود پاکت بیش از ۴ میلی‌متر ($۱۵٫۲۶۹ = \hat{\theta}_3 \leqslant y^*_{ij}$) تغییر دهد.

همان‌طور که پیش از این بیان شد، در مدل پیشنهادی ما، مقدار ضریب همبستگی بین پاسخ‌های فک بالا و پایین را می‌توان به‌وسیله‌ی براورد پارامتر همبستگی مدل (v) محاسبه نمود. با توجه به فرمول ارائه شده در بخش ۳٫۲ و مقدار براورد شده‌ی $\hat{v} = ۰/۴۴۷$، ضریب همبستگی ρ را می‌توان به‌صورت زیر براورد نمود:

$$\hat{\rho} = \frac{\zeta(۲, ۰٫۴۴۷)}{\zeta(۲, ۰٫۴۴۷) + \frac{۱}{۶}\pi^۲};$$

یعنی براورد ضریب همبستگی بین پاسخ‌های مشاهده شده در مدل نامتقارن، در حدود ۰٫۷۸۴ بوده است. همچنین، با توجه به مدل کیم، ضریب همبستگی بین پاسخ‌ها را می‌توان مستقیماً با براورد کردن ضریب همبستگی توزیع نرمال دومتغیره (توزیع نهانی) به دست آورد. این براورد برای داده‌های مورد مطالعه، در حدود ۰٫۸۰ محاسبه شده است، که تفاوت چندانی با براورد حاصل از مدل نامتقارن ندارد. توجه داریم که مقدار ضریب همبستگیِ به دست آمده از مدل نامتقارن، به ضریب همبستگی اسپیرمن ($r =$ ۰٫۷۴۱)

© ۱۳۸۳ پژوهشکده‌ی آمار

نزدیک‌تر است تا به براورد حاصل از مدل پروبیت تجمعی دومتغیره.

۵ بحث

هدف ما از ارائه‌ی این مقاله، معرفی روش‌هایی برای مدل‌بندی پاسخ‌های جفت‌شده‌ی ترتیبی بود. بدین منظور، بر مدل‌هایی که یک توزیع دومتغیره‌ی پیوسته‌ی نهانی را به‌عنوان تابع پیوند مورد استفاده قرار می‌دهند، تأکید بیش‌تری شد. اوّلین پرسش در این جهت ممکن است بدین صورت مطرح شود که آیا این مدل‌ها با توجه به پیچیده و پردردسر بودنِ برازاندن‌شان، واقعاً اطلاعاتی بیش از نتایج حاصل از برازش مدل‌های حاشیه‌ای یک‌متغیره به دست می‌دهند؟ پاسخ را به این صورت می‌توان بیان کرد:

۱. هنگامی که مدل‌های حاشیه‌ای را مورد بررسی قرار می‌دهیم، ممکن است رابطه‌ی متغیرهای کمکی مختص فرد یا اندام با پاسخ سمت راست (یا بالا) معنادار و با پاسخ سمت چپ (یا پایین) غیر معنادار باشد. چنین نتایجی ممکن است ما را در استنباط‌های نهایی دچار اشکال کند.

۲. مدل‌های دومتغیره این امکان را برای ما فراهم می‌سازند که براوردی صریح از ضریب همبستگی بین متغیرهای پاسخ در حضور متغیرهای کمکی مختلف به دست آوریم. این امکان در مدل‌های یک‌متغیره وجود ندارد.

۳. در مدل‌های دومتغیره قادریم اثر عوامل خطر مختص یک اندام را در حالی که همبستگی بین این اندام و اندام دوم به‌طور همزمان در مدل احتساب شده است، مورد بررسی قرار دهیم. مدل‌های یک‌متغیره‌ی حاشیه‌ای چنین امکانی را در اختیار ما قرار نمی‌دهند.

با در نظرگرفتن موارد فوق، هنگامی که نمونه‌ای با اندازه‌ی بزرگ در اختیار داریم، بهتر است مزایای مدل‌های دومتغیره را در مقابل عیب بزرگ آن‌ها (پیچیده بودن برازش) قرار داده، در مورد انتخاب مدل مناسب تصمیم‌گیری کنیم.

مطلب دیگری که در این بخش باید به آن اشاره شود، این است که مدل‌های دومتغیره‌ی توصیف‌شده در بخش سوم را به‌سادگی می‌توان برای کاربردهای دیگر تعمیم داد. به‌عنوان مثال، می‌توان پارامترهای رگرسیونی مختلفی برای حاشیه‌ها در نظرگرفت. در این حالت، کافی است ماتریس طرح را به‌صورت زیر تغییر داده:

$$\boldsymbol{X}_i^{'} = \begin{pmatrix} \boldsymbol{x}_i^{'} & \boldsymbol{x}_{i1}^{'} & \mathbf{0} \\ \boldsymbol{x}_i^{'} & \mathbf{0} & \boldsymbol{x}_{i2}^{'} \end{pmatrix},$$

© ۱۳۸۳ پژوهشکده‌ی آمار

پارامترهای رگرسیونی متناظر را به‌صورت $(\boldsymbol{\delta}, \varepsilon_1, \varepsilon_2)$ در نظر بگیریم، که در آن، ε_1 پارامترهای رگرسیونی مربوط به متغیرهای کمکیِ به دست آمده از سمت راست، و ε_2 پارامترهای رگرسیونی مرتبط با متغیرهای کمکیِ به دست آمده از سمت چپ هستند. همچنین در صورتی که داده‌ها ایجاب کند، قادریم نقاط برش متفاوتی برای حاشیه‌ها در نظر بگیریم. اگر افراد مورد مطالعه بر حسب دو مقیاس ترتیبی متفاوت دسته‌بندی شده باشند، در نظرگرفتن نقاط برش مختلف برای حاشیه‌ها ضروری به نظر می‌رسد. در مورد داده‌های CPI، از آن‌جایی که دسته‌بندی متغیر پاسخ برای هر دو حاشیه کاملاً یکسان بوده است، چنین تعمیمی منطقی نخواهد بود.

نکته‌ی قابل ذکر دیگر، انتخاب حدس‌های اوّلیه‌ی مناسب برای آغاز فرایند براورد پارامترهای رگرسیونی است. اگر حدس‌های اوّلیه به‌طور نادرست و دور از مقادیر واقعی انتخاب شوند، ممکن است روش‌های تکراری در ماکسیمم‌های محلی (local) گرفتار شوند و براوردهای گمراه‌کننده‌ای نتیجه دهند. بر اساس تجربه، ما از حدس‌های اوّلیه‌ی $\hat{v} = 1$، $\hat{\delta} = 0$، و $\hat{\varepsilon} = 0$ برای آغاز این فرایند استفاده کردیم. همچنین برای انتخاب حدس‌های اوّلیه در مورد نقاط برش مدل، می‌توان از براورد GEE یا نتایجِ به دست آمده از برازش مدل‌های حاشیه‌ای استفاده کرد. در این‌جا ذکر این نکته ضروری است که در صورت طراحی یک برنامه‌ی رایانه‌ای مناسب، براوردهای نهایی با تغییرات جزئی در حدس‌های اوّلیه، تغییر نخواهند کرد. به‌عنوان مثال، در برنامه‌ی رایانه‌ای طراحی شده توسط ما و کیم، با انتخاب حدس‌های اوّلیه‌ی مناسب و ایجاد تغییرات جزئی در آن‌ها، مقادیر براوردهای نهایی دستخوش تغییر نمی‌شوند. انتخاب حدس‌های اوّلیه‌ی اندکی نامناسب، فقط سبب افزایش تعداد تکرارهای لازم برای رسیدن به همگرایی نهایی، و در نتیجه موجب بالا رفتن مدت زمان پردازش رایانه‌ای خواهد شد.

مطلب دیگری که در این قسمت باید به آن اشاره شود این است که مدل‌های مبتنی بر توزیع‌های نهانی توأم، چه برتری‌ای نسبت به مدل‌های مبتنی بر براوردهای GEE دارند. همان‌طورکه از نتایج مندرج در جدول ۲ مشخص است، در مدل‌های مبتنی بر توزیع‌های نهانی دومتغیره، همبستگی بین مشاهدات پاسخ به‌عنوان یک پارامتر مهم در مدل وارد می‌شود و در پایان، براوردی صریح از این پارامتر در حضور متغیرهای کمکی مختلف به دست می‌آید. اما در روش GEE، همبستگی بین مشاهدات پاسخ به‌عنوان یک پارامتر مزاحم در نظرگرفته می‌شود و به‌علتِ به کار بردن توزیع‌هایی حاشیه‌ای به‌جای توزیع توأم متغیرهای پاسخ، در حین فرایند مدل‌سازی، براورد صریحی از این همبستگی به دست نمی‌آید. این مطلب می‌تواند موجب برازش ضعیف‌تر مدل‌های مبتنی بر روش GEE نسبت به مدل‌هایی با براورد حد اکثر درست‌نمایی شود.

به‌عنوان توصیه‌ی نهایی، ذکر این نکته الزامی است که مدل‌های مبتنی بر توزیع‌های نهانی که در این مقاله به آن‌ها اشاره شد، بهتر است هنگامی مورد استفاده قرار گیرند که پاسخی دومتغیره و ترتیبی با همبستگی نسبتاً بالا در اختیار داشته باشیم. برای تصمیم‌گیری در مورد انتخاب مدل ما یا مدل پروبیت تجمعیِ

© ۱۳۸۳ پژوهشکده‌ی آمار

کیم، می‌توان از برازش‌های حاشیه‌ای مدل‌های یک‌متغیره استفاده کرد. اگر مدل‌های پروبیت یا لوژستیک تجمعی برای پاسخ‌های حاشیه‌ای، برازش‌های بهتری نسبت به مدل‌های غیر متقارن مانند log-log به دست دهند، انتظار داریم که مدل کیم بهتر از مدل ما عمل کند. در غیر این صورت، مدل ارائه شده از سوی ما احتمالاً (و نه لزوماً) جایگزینی مناسب برای مدل کیم خواهد بود.

مرجع‌ها

[1] Ananth, C.V.; Kleinbaum, D.G. (1997). Regression models for ordinal responses: a review of methods and applications. *Int. J. Epidemiol.* **26**, 1323-1333.

[2] Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York.

[3] Biswas, A.; Das, K. (2002). A Bayesian analysis of bivariate ordinal data: Wisconsin epidemiologic study of diabetic retinopathy revisited. *Statist. Med.* **21**, 549-559.

[4] Dos Santos, D.M.; Berridge, D.M. (2000). A continuation ratio random effects model for repeated ordinal responses. *Statist. Med.* **19**, 3377-3388.

[5] Gange, S.J.; Linton, K.L.; Scott, A.J.; DeMets, D.L.; Klein, R. (1995). A comparison of methods for correlated ordinal measures with ophthalmic applications. *Statist. Med.* **14**, 1961-1974.

[6] Gumbel, E.J. (1961). Bivariate logistic distributions. *J. Amer. Statist. Assoc.* **56**, 335-349.

[7] Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statist. Med.* **14**, 1341-1352.

[8] Liang, K.Y.; Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

[9] Lipsitz, S.R.; Kim, K.; Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statist. Med.* **13,** 1149-1163.

[10] Lipsitz, S.R.; Laird, N.M.; Harrington, D.P. (1990). Maximum likelihood regression methods for paired binary data. *Statist. Med.* **9**, 1517-1525.

[11] McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc.* **B 42**, 109-142.

[12] Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics* **40**, 1025-1035.

[13] Satterthwaite, S.P.; Hutchinson, T.P. (1978). A generalisation of Gumbel's bivariate logistic distribution. *Metrika* **25**, 163-170.

[14] Williamson, J.M.; Kim, K.; Lipsitz, S.R. (1995). Analyzing bivariate ordinal data using a global odds ratio. *J. Amer. Statist. Assoc.* **90**, 1432-1437.

© ۱۳۸۳ پژوهشکده‌ی آمار

[15] Williamson, J.M.; Lipsitz, S.R.; Kim, K.M. (1999). GEECAT and GEEGOR: computer programs for the analysis of correlated categorical response data. *Comput. Methods Programs Biomed.* **58**, 25-34.

دریافت: ۱ تیر ۱۳۸۳
آخرین اصلاح: ۲۷ آبان ۱۳۸۳
انتشار: ۲۰ دی ۱۳۸۳

انوشیروان کاظم‌نژاد
گروه آمار زیستی، دانشکده‌ی علوم پزشکی،
دانشگاه تربیت مدرس،
پل نصر، بزرگراه جلال آل احمد،
تهران، ایران.
پیام‌نگار: *kazem_an@modares.ac.ir*

فرید زایری
گروه آمار زیستی، دانشکده‌ی علوم پزشکی،
دانشگاه تربیت مدرس،
پل نصر، بزرگراه جلال آل احمد،
تهران، ایران.
پیام‌نگار: *fzayeri@yahoo.com*

© ۱۳۸۳ پژوهشکده‌ی آمار

Modeling Paired Ordinal Response Data

Anoshiravan Kazemnejad and Farid Zayeri

Tarbiat Modarres University

Abstract. About 25 years ago, McCullagh proposed a method for modeling univariate ordinal responses. After publishing this paper, other statisticians gradually extended his method, such that we are now able to use more complicated but efficient methods to analyze correlated multivariate ordinal data, and model the relationship between those responses and host of covariates. In this paper, we aim to present the recent progressions in modeling ordinal response data, especially in bivariate ordinal responses that arise from medical studies relating to paired organs such as ophthalmology, otology, nephrology, etc. Additionally, we present a new model for analyzing correlated ordinal response data. This model is an appropriate alternative for bivariate cumulative probit regression model, when joint distribution of response data is not symmetric. Finally, as an applied example, we analyze the obtained data from an epidemiologic study relating to periodontal status among high school students in Tehran using this method and compare the results with the similar models.
© 2004 Statistical Research Center. All rights reserved.

Keywords. correlated ordinal responses; bivariate latent distribution; generalized estimating equations; generalized linear models.