تحلیل پاسخ‌های آمیخته‌ی گسسته و پیوسته‌ی رده‌بندی شده

احسان بهرامی‌سامانی[۱] و حمیدرضا عالی‌پور

بخش آمار، دانشگاه شهید بهشتی

تاریخ دریافت: ۹۲/۴/۳۱ تاریخ پذیرش: ۹۳/۳/۲۰

چکیده: در این مقاله، هدف ما رده‌بندی پاسخ‌هایی است که به‌صورت آمیخته‌ای از پاسخ‌های گسسته و پیوسته هستند. برای این‌کار، ابتدا باید تابع توزیع توأم چنین پاسخ‌هایی را داشته باشیم. بنابراین، مدل مکانی عام جدید را برای دستیابی به تابع توزیع توأم پاسخ‌های آمیخته‌ی گسسته و پیوسته معرفی کرده‌ایم و به‌صورت اجمالی با مدل مکانی عام دِ لئون و کریر [۱] مورد مقایسه قرار داده‌ایم. رویکرد مورد استفاده برای رده‌بندی چنین پاسخ‌هایی، استفاده از قاعده‌ی رده‌بندی ولچ [۲] به‌همراه ارائه‌ی احتمال‌های بد-رده‌بندی و نرخ واقعی خطا برای این مدل است. همچنین نتایج رده‌بندی مدل مکانی عام جدید با نتایج رده‌بندی مدل مکانی عام مورد مقایسه قرار گرفته است [۱]. لازم به ذکر است قواعد رده‌بندی نمونه‌ای برای مدل مکانی عام جدید با استفاده از برآوردهای به‌دست آمده از طریق ماکسیمم‌سازی تابع درستنمایی مطرح می‌شود. در پایان، برای نشان دادن قابلیت قواعد به‌دست آمده، مطالعه‌ی شبیه‌سازی و تحلیل داده‌های واقعی با استفاده از مدل مکانی عام جدید انجام شده است.

واژه‌های کلیدی: قواعد رده‌بندی، مدل مکانی عام، پاسخ‌های آمیخته‌ی پیوسته و گسسته، تابع درستنمایی، احتمال بد-رده‌بندی.

رده بندی ریاضی: ۶۲J۰۵، ۶۲J۱۲.

۱- آدرس الکترونیکی نویسنده مسئول مقاله: احسان بهرامی‌سامانی e_bahrami@sbu.ac.ir

۱- مقدمه

تفکیک و رده‌بندی داده‌ها، این‌که کدام داده به کدام جامعه تعلق دارد، همواره از مسائل مورد توجه در زمینه‌ی علوم پزشکی و سایر علوم بوده است. در این بین مواردی نیز پیش می‌آید که در آن پاسخ‌های به‌دست آمده در زمینه‌ای خاص، صرفاً از یک نوع، یعنی فقط پیوسته یا فقط گسسته، نیستند بلکه آمیخته‌ای از هر دو نوع، یعنی متشکل از پاسخ‌های گسسته و پاسخ‌های پیوسته هستند. مسئله وقتی جذاب‌تر می‌شود که بدانیم بین پاسخ‌ها همبستگی نیز وجود دارد. بنابراین، دیگر استفاده از روش‌های استاندارد قبلی کاری نادرست بوده و نتایج به‌دست آمده از آن‌ها غیر از آن نتایجی است که باید به‌دست آیند. این نوع پاسخ‌ها خود به انواع مختلفی تقسیم‌بندی می‌شوند که از آن‌جمله می‌توان به پاسخ‌های آمیخته‌ی پیوسته و اسمی، پاسخ‌های پیوسته، اسمی و ترتیبی و غیره اشاره کرد. آنچه در این بین پژوهش‌گر را به چالش می‌کشاند، مسئله‌ی تعیین توزیع توأم این نوع پاسخ‌ها برای دستیابی هر چه آسان‌تر به محاسبه‌ی مقدار همبستگی و نیز سایر خصیصه‌های مورد نیاز وی در این‌گونه پاسخ‌ها است. رویکردهای گوناگونی برای این منظور و توسط افراد مختلف ارائه شده است ولی اکثر آن‌ها دارای انعطاف‌پذیری کافی نیستند و شکل توزیعی استانداردی ندارند. تیت از جمله‌ی پیشگامان در این زمینه است [۳].

یکی از پیشنهادهای قابل‌توجه در این زمینه، رویکرد مستقیم توزیع توأم پاسخ‌های آمیخته است. به این صورت که توزیع توأم این نوع پاسخ‌ها، از طریق تجزیه به توزیع حاشیه‌ای مجموعه‌ای از پاسخ‌ها و توزیع شرطی سایر پاسخ‌ها به شرط این مجموعه به‌دست می‌آید. دو نوع مدل می‌توان برای این روش در نظر گرفت. در مدل اول که به مدل مکانی عام (GLOM)[1] معروف است، توزیع توأم پاسخ آمیخته از طریق تجزیه به توزیع حاشیه‌ای پاسخ گسسته و توزیع شرطی پاسخ پیوسته به شرط پاسخ گسسته مهیا می‌شود و فرض بر این است که پاسخ‌های پیوسته به شرط پاسخ‌های گسسته دارای توزیع نرمال چند متغیره با میانگین‌های متفاوت و ماتریس کوواریانس یکسان و پاسخ‌های گسسته دارای توزیع چندجمله‌ای حاشیه‌ای باشند. در مدل دیگر که به مدل پیوسته‌ی گروه‌بندی شده‌ی شرطی (CGCM)[2] معروف است، توزیع توأم با استفاده از تجزیه به توزیع حاشیه‌ای پاسخ پیوسته و توزیع شرطی پاسخ گسسته به شرط پاسخ پیوسته به‌دست می‌آید. در زمینه‌ی

1- General Location Model
2- Conditional Grouped Continuous Model

بررسی این نوع مدل، می‌توان به کاکس و ویرموس اشاره کرد [۴]. همچنین در این زمینه، کاتالانو و رایان در [۵] مدل داده‌های آمیخته بر پایه‌ی متغیر پنهان را مورد بررسی قرار داده‌اند. از سوی دیگر، دِ لئون و کریر در [۱] ، یک مدل مکانی عام آمیخته با شامل متغیرهای اسمی و پاسخ های پیوسته و ترتیبی معرفی نمودند به طوری که دراین مدل، جدول پیشایندی از تقاطع سطوح متغیرهای اسمی حاصل می شود و پاسخ‌های پیوسته و ترتیبی درون خانه‌های این جدول قرار می‌گیرند و توزیع توأم پاسخ‌های آمیخته‌ی پیوسته، ترتیبی و اسمی از طریق تجزیه به توزیع پاسخ‌های پیوسته و ترتیبی به شرط متغیرهای اسمی و توزیع حاشیه‌ای متغیرهای اسمی به دست می آید.

در ادامه دِ لئون و همکاران در [۶] برای اولین بار به بررسی رده بندی داده ها ی آمیخته در مدل دِ لئون و کریر در [۱] پرداختند. اما آنچه در این مقاله مطرح می‌شود، پیشنهاد مدل جدیدی است که در آن جدول پیشایندی از تقاطع هم‌زمان سطوح پاسخ‌های اسمی و ترتیبی تشکیل می‌شود و این پاسخ‌های پیوسته هستند که درون خانه‌های جدول پیشایندی قرار می‌گیرند. تابع توزیع توأم نیز از طریق تجزیه به تابع توزیع شرطی پاسخ‌های پیوسته به شرط پاسخ‌های ترتیبی، تابع توزیع شرطی پاسخ‌های ترتیبی به شرط پاسخ‌های اسمی و تابع توزیع حاشیه‌ای پاسخ‌های اسمی به‌دست می‌آید. رویکرد رده‌بندی در این مقاله نیز قاعده‌ی رده‌بندی ولچ [۲] خواهد بود و همان‌طور که اشاره خواهد شد، رده‌بندی این نوع پاسخ‌ها با استفاده از مدل جدید، نتایج بهتری در مقایسه با مدل دِ لئون و کریر [۱] ارائه می‌کند. برای این منظور، ابتدا احتمال‌های بد-رده‌بندی حاصل از مدل مکانی عام جدید با مورد مشابه برای مدل دِ لئون و کریر [۱] با استفاده از مطالعه‌ی شبیه‌سازی مورد مقایسه قرارگرفته و سپس برای تشریح سودمندی رده‌بندی مدل مکانی عام جدید، مطالعه‌ای روی بیماری کبد چرب و رده‌بندی بیماران صورت گرفته است.

این مقاله به این صورت سازماندهی شده است که در بخش دوم، مروری کوتاه بر مدل مکانی عام دِ لئون و کریر [۱] داشته، سپس به معرفی مدل مکانی عام جدید و نیز ارائه تابع درستنمایی و قواعد رده‌بندی این مدل می‌پردازیم. در بخش سوم برای مقایسه‌ی مدل مکانی عام جدید و مدل مکانی عام دِ لئون و کریر [۱] از یک مطالعه‌ی شبیه‌سازی استفاده‌شده و در نهایت در بخش چهارم، قواعد رده‌بندی مربوط به مدل مکانی عام دِ لئون و کریر [۱] و مدل مکانی عام جدید بر روی بیماری کبد چرب بیماران بیمارستان طالقانی تهران پیاده‌سازی شده است.

۲- مدل‌های مربوط به پاسخ‌های آمیخته‌ی همبسته‌ی پیوسته و گسسته

در پژوهش‌های آماری مواردی اتفاق می‌افتد که پژوهش‌گر با دو نوع پاسخ پیوسته و گسسته به‌صورت هم‌زمان و در یک مدل روبرو می‌شود و پیچیدگی چنین مدل‌هایی وقتی ملموس‌تر می‌شود که بین چنین پاسخ‌هایی همبستگی نیز وجود داشته باشد. در این‌گونه موارد، به تنهایی استفاده از مدل‌های پاسخ‌های پیوسته یا حتی مدل‌های پاسخ‌های گسسته که در ادبیات آماری به وفور یافت می‌شوند، عملی ناشیانه است و نتایج اشتباهی را در بر دارد. آنچه که ماهیت این نوع پاسخ‌ها ایجاب می‌کند، یافتن روش مناسب برای تحلیل چنین پاسخ‌های آمیخته است. در سال‌های اخیر و با مطرح‌شدن تحلیل داده‌های آمیخته، فعالیت‌های پژوهشی گسترده‌ای در این زمینه صورت گرفته است و ماحصل آن بیان سه نوع رویکرد مختلف در حالت‌ها و مدل‌های مختلف است. لازم به ذکر است که در مدل‌هایی با پاسخ‌های آمیخته‌ی گسسته و پیوسته، بیش‌ترین چالش فراروی پژوهش‌گران آماری، نحوه‌ی به‌دست آوردن توزیع توأم این نوع پاسخ‌هاست. اولین رویکردی که برای تحلیل پاسخ‌های توأم پیوسته و گسسته استفاده می‌شود و از شیوه‌ای مستقیم برای یافتن تابع توزیع توأم این نوع پاسخ‌ها استفاده می‌کند، مدل تجزیه به عامل‌ها[1] است. این رویکرد نیز خود به دو روش مدل مکانی عام (GLOM) و مدل پیوسته‌ی گروه‌بندی شده‌ی شرطی (CGCM) تقسیم می‌شود. در مدل مکانی عام، تابع توزیع توأم پاسخ‌های آمیخته‌ی گسسته و پیوسته از طریق تجزیه به توزیع حاشیه‌ای برای پاسخ‌های گسسته و توزیع شرطی پاسخ‌های پیوسته به شرط پاسخ‌های گسسته به‌دست می‌آید. در مدل پیوسته‌ی گروه‌بندی شده‌ی شرطی، توزیع توأم پاسخ‌های آمیخته‌ی گسسته و پیوسته از طریق تجزیه‌ی آن به توزیع حاشیه‌ای برای پاسخ‌های پیوسته و توزیع شرطی برای پاسخ‌های گسسته به شرط پاسخ‌های پیوسته به‌دست می‌آید. رویکرد دوم که به رویکرد غیرمستقیم شهرت دارد و به دو دسته‌ی مفصل‌ها [۷] و اثرهای تصادفی [۸] تقسیم می‌شود به همراه رویکرد سوم که شامل مدل‌های پلاکت-دیل [۹] و توزیع‌های نمایی درجه‌ی دو [۱۰] است، در این مقاله مورد مطالعه قرار نمی‌گیرند.

1- Factorization model

۲-۱- مدل مکانی عام با پاسخ‌های آمیخته‌ی پیوسته، ترتیبی و اسمی رده‌بندی شده

در این بخش، ابتدا مدل مکانی عام با پاسخ‌های آمیخته‌ی پیوسته و اسمی، آمیخته‌ی پیوسته و ترتیبی و مدل مکانی عام با پاسخ‌های آمیخته‌ی پیوسته، ترتیبی و اسمی معرفی می‌گردد [۱]. سپس مدل مکانی عام جدید با پاسخ‌های آمیخته‌ی پیوسته، ترتیبی و اسمی معرفی و مسئله‌ی رده‌بندی روی این مدل بیان می‌گردد. لازم به ذکر است مسئله‌ی رده‌بندی روی مدل مکانی عام جدید، توسط محققان دیگر مورد بررسی قرار نگرفته است.

۲- ۱- ۱- مقدمه‌ای بر مدل مکانی عام

فرض کنیم $\mathbf{u}$ و $\mathbf{n}$ به ترتیب به‌عنوان بردار پاسخ‌های پیوسته و اسمی باشند. از تقاطع سطوح متغیرهای اسمی، جدول پیشایندی D خانه به‌وجود می‌آید و بردار متغیرهای پیوسته درون خانه‌های این جدول قرار می‌گیرند. دِ لئون و کریر [۱] با تعمیم این مدل، مدل جدیدی را برای سه متغیر پیوسته، ترتیبی و اسمی آمیخته ارائه کرده‌اند. در این مدل همانند مدل قبلی، متغیرهای اسمی تشکیل جدول پیشایندی می‌دهند و متغیرهای پیوسته به‌همراه متغیرهای ترتیبی بر پایه‌ی متغیر پنهان درون خانه‌های این جدول قرار می‌گیرند.

مثال ۱. متغیر اسمی N_1 (۱ یا $N_1 = 0$) و متغیر پیوسته‌ی U را در نظر بگیریم. با توجه به دو سطحی بودن متغیر اسمی، جدول پیشایندی با دو خانه تشکیل می‌شود که متغیر پیوسته درون خانه‌های این جدول قرار می‌گیرد و با فرض ثابت بودن واریانس درون این دو خانه، میانگین آن از خانه‌ای به خانه‌ی دیگر تغییر می‌کند. این میانگین‌ها در شکل ۱ با دو نماد μ_1 و μ_2 نشان داده شده است.

N_1	
۰	۱
μ_1	μ_2

شکل ۱: مدل مکانی عام دِ لئون و کریر [۱] برای پاسخ‌های پیوسته و اسمی

مثال ۲. فرض کنیم متغیر پیوسته‌ی U و متغیرهای ترتیبی O_1 و O_2 به‌ترتیب با دو و سه سطح موجود باشند.

$$O_۲=\begin{cases}۱ & U_۲^*\le\theta_۱\\ ۲ & \theta_۱<U_۲^*\le\theta_۲\\ ۳ & U_۲^*>\theta_۲\end{cases} \quad \text{و} \quad O_۱=\begin{cases}۱ & U_۱^*<\lambda\\ ۲ & U_۱^*\ge\lambda\end{cases}$$

که در آن $U_۱^*$ و $U_۲^*$ به ترتیب متغیرهای پنهان مربوط به متغیرهای ترتیبی $O_۱$ و $O_۲$ و λ، $\theta_۱$ و $\theta_۲$ نقاط آستانه‌ای مربوط به متغیرهای ترتیبی هستند. جدول پیشایندی حاصل از تقاطع سطوح متغیرهای ترتیبی، دارای شش خانه خواهد بود که متغیر پیوسته درون خانه‌های این جدول قرار می‌گیرد، منتهی میانگین این متغیر با فرض ثابت بودن واریانس آن از خانه‌ای به خانه‌ی دیگر فرق می‌کند که در شکل ۲ با نماد μ_j ($j=۱,...,۶$) نشان داده شده است.

$U_۱^*\ge\lambda$	$U_۱^*<\lambda$	
$\mu_۲$	$\mu_۱$	$U_۲^*\le\theta_۱$
$\mu_۴$	$\mu_۳$	$\theta_۱<U_۲^*\le\theta_۲$
$\mu_۶$	$\mu_۵$	$U_۲^*>\theta_۲$

شکل ۲: مدل مکانی عام دِ لئون و کریر [۱] برای پاسخ‌های پیوسته و ترتیبی

مثال ۳. فرض کنیم متغیر پیوسته‌ی $U_۱$ و $U_۲$، متغیرهای ترتیبی $O_۱$ و $O_۲$ (متغیرهای پنهان متناظر با آن‌ها به ترتیب $U_۱^*$ و $U_۲^*$ هستند) به ترتیب با دو و سه سطح (مانند مثال ۲) و متغیرهای اسمی $N_۱$ و $N_۲$ هر کدام با دو سطح موجود باشند.

$N_۱$					
۱		۰			
$U_۱$	$U_۱^*$	$U_۱$	$U_۱^*$		
$U_۲$	$U_۲^*$	$U_۲$	$U_۲^*$	۰	$N_۲$
$U_۱$	$U_۱^*$	$U_۱$	$U_۱^*$	۱	
$U_۲$	$U_۲^*$	$U_۲$	$U_۲^*$		

شکل ۳: مدل مکانی عام دِ لئون و کریر [۱] برای پاسخ‌های پیوسته، اسمی و ترتیبی

جدول پیشایندی حاصل از تقاطع سطوح متغیرهای اسمی با چهارخانه تشکیل خواهد شد که متغیرهای پیوسته و ترتیبی درون خانه‌های آن قرار خواهند گرفت.

۲- ۱- ۲- مدل و تابع درستنمایی مدل مکانی عام جدید

در مدل جدید، در حالت وجود متغیرهای اسمی، ترتیبی و پیوسته جدول پیشایندی برخلاف مدل دِ لئون و کریر [۱] که از تقاطع متغیرهای اسمی تشکیل می‌شد، هم‌زمان از تقاطع سطوح متغیرهای اسمی و ترتیبی تشکیل می‌شود و متغیرهای پیوسته درون خانه‌های جدول پیشایندی قرار می‌گیرند. این مطلب با فرض وجود متغیرهای مثال قبلی در شکل ۴ نشان داده شده است.

		$N_۲$						
		۰			۱			
		$U_۲^* \le \theta_۱$	$\theta_۱ < U_۲^* \le \theta_۲$	$U_۲^* > \theta_۲$	$U_۲^* \le \theta_۱$	$\theta_۱ < U_۲^* \le \theta_۲$	$U_۲^* > \theta_۲$	
$N_۱$	۰	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱^* < \lambda$
		$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱^* \ge \lambda$
	۱	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱^* < \lambda$
		$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱$ $U_۲$	$U_۱^* \ge \lambda$

شکل ۴: مدل مکانی عام جدید برای پاسخ‌های پیوسته، اسمی و ترتیبی

نکته‌ی حائز اهمیت این است که با فرض وجود بردار پاسخ‌های آمیخته‌ی اسمی $(\mathbf{n})$، ترتیبی $(\mathbf{o})$ و پیوسته $(\mathbf{u})$، تابع چگالی توأم آن‌ها در مدل مکانی عام دِ لئون و کریر [۱] به‌صورت $f(\mathbf{u},\mathbf{o}\,|\,\mathbf{n})f(\mathbf{n})$ و در مدل مکانی عام جدید به‌صورت $f(\mathbf{u}\,|\,\mathbf{n},\mathbf{o})f(\mathbf{n},\mathbf{o})$ تجزیه می‌شود و تعداد خانه‌های جدول مدل مکانی عام جدید بیشتر از تعداد خانه‌های جدول پیشایندی مدل مکانی عام دِ لئون و کریر [۱] است.

در حالت کلی، فرض کنیم $\mathbf{n}$ و $\mathbf{o}$ به ترتیب بردارهایی $S\times ۱$ و $Q\times ۱$ از متغیرهای اسمی و ترتیبی هستند، که jامین مؤلفه‌های آن‌ها دارای d_j و d_i^* حالت ممکن می‌باشند. زوجِ $(\mathbf{n},\mathbf{o})$ تشکیل یک جدول پیشایندی با $D=\prod_{i=۱}^{Q}\prod_{j=۱}^{S} d_i^* d_j$ حالت ممکن یا به عبارتی D خانه را می‌دهند. بردار اسمی مانند $\mathbf{x}=(X_۱,...,X_D)$ را می‌توان به گونه‌ای تعریف کرد که اگر $(\mathbf{n},\mathbf{o})$ در حالت d، $d=۱,۲,...,D$، قرار گیرد، X_d برابر یک و در غیر این صورت برابر صفر باشد.

بنابراین برای فرد خاصی، $\mathbf{x}$ فقط شامل یک عدد ۱ خواهد بود. تحت مدل مکانی عام، $\mathbf{x}$ دارای توزیع چندجمله‌ای است؛

$$\mathbf{x}\ |\boldsymbol{\pi}\sim\prod_{d=۱}^{D}\pi_d^{X_d} \qquad (۱)$$

که در آن $\boldsymbol{\pi}=(\pi_۱,...,\pi_D)^T$ برداری از احتمال‌های مربوط به خانه‌های جدول پیشایندی است. متغیرهای نشانگر را برای فرد kـام ($k=۱,...,\mathrm{n}$) به‌صورت زیر تعریف می‌کنیم:

$$I_{u_k}=\begin{cases}۱ & U_k\le u_k\\ ۰ & o.w.\end{cases},\quad I_{u_k^*}=\begin{cases}۱ & \lambda_k<U_k^*\le\lambda_{k+۱}\\ ۰ & o.w.\end{cases},\quad I_{n_k}=\begin{cases}۱ & N_k=n_k\\ ۰ & o.w.\end{cases}$$

که در آن λ_k نقطه‌ی برشی برای متغیر پنهان U_k^* است. ارتباط بین متغیر پنهان و متغیر ترتیبی به وسیله‌ی مدل آستانه‌ای تعریف می‌شود، یعنی،

$$O_q=\lambda_q^{l_q}\ \Leftrightarrow\ \lambda_q^{l_q-۱}<U_q^*\le\lambda_q^{l_q},\qquad q=۱,...,Q$$

که در آن $\lambda_q^{l_q}$ نقاط آستانه‌ای نام دارند و مجموعه‌ی $\left\{\lambda_q^{۱},...,\lambda_q^{l_q}\right\}$ نقاط نامعلوم هستند. همچنین $\lambda_q^{۰}=-\infty$ و $\lambda_q^{l_q+۱}=+\infty$. فرض کنیم متغیرهای ترتیبی $O_۱$ و $O_۲$ و متغیرهای پنهان $U_۱^*$ و $U_۲^*$ به ترتیب متناظر با $O_۱$ و $O_۲$ و متغیر پیوسته U_k موجود باشند. تابع درستنمایی مدل عبارت است از:

$$
\begin{aligned}
&L(\boldsymbol{\eta}\,|\,\mathbf{u},O_1,O_2,\mathbf{n})\\
&=\prod_{k=1}^{n}\frac{\partial}{\partial u_k}\Big\{\Pr(U_k\le u_k\,|\,\lambda_{k1}^{l_{k1}-1}<U_{k1}^{*}\le\lambda_{k1}^{l_{k1}},\lambda_{k2}^{l_{k2}-1}<U_{k2}^{*}\le\lambda_{k2}^{l_{k2}},n_k,\boldsymbol{\Sigma})\Big\}\\
&\quad\times\Pr(\lambda_{k2}^{l_{k2}-1}<U_{k2}^{*}\le\lambda_{k2}^{l_{k2}}\,|\,\lambda_{k1}^{l_{k1}-1}<U_{k1}^{*}\le\lambda_{k1}^{l_{k1}},n_k,\lambda_q^{1},\lambda_q^{2},...,\lambda_q^{l_q})\\
&\quad\times\Pr(\lambda_{k1}^{l_{k1}-1}<U_{k1}^{*}\le\lambda_{k1}^{l_{k1}}\,|\,n_k,\lambda_q^{1},\lambda_q^{2},...,\lambda_q^{l_q})\times\Pr(N_k=n_k)
\end{aligned}
$$

که در آن $\boldsymbol{\Sigma}$ ماتریس کوواریانس است. همچنین احتمال مربوط برای خانه‌ی dام جدول پیشایندی برای $d=1,2,...,D$ را که در آن D تعداد کل خانه‌ها است، به‌صورت زیر تعریف می‌کنیم:

$$
\pi_d=\Pr\left(\lambda_{k1}^{l_{k1}-1}<U_{k1}^{*}\le\lambda_{k1}^{l_{k1}},\lambda_{k2}^{l_{k2}-1}<U_{k2}^{*}\le\lambda_{k2}^{l_{k2}}\,|\,N_k=n_k\right)\times\Pr(N_k=n_k)\,.
$$

با استفاده از تقریب توزیع نرمال چند متغیره که توسط جوی [۱۱] ارائه شده است، تابع درستنمایی به‌صورت زیر بیان می‌شود:

$$
\begin{aligned}
L(\boldsymbol{\eta}\,|\,\mathbf{u},O_1,O_2,\mathbf{n})\cong\prod_{k=1}^{n}\Bigg[&(2\pi)^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{u}_k-\boldsymbol{\mu}_d)^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{u}_k-\boldsymbol{\mu}_d)\right\}\\
&+\frac{\partial}{\partial u_k}\boldsymbol{\omega}_{21}^{T}\boldsymbol{\omega}_{11}^{-1}\left[1-E\left(I_{u_{k1}^{*}}\right),1-E\left(I_{u_{k2}^{*}}\right),1-E\left(I_{n_k}\right)\right]^{T}\pi_d\Bigg]
\end{aligned}
\qquad(2)
$$

که در آن

$$
E\left(I_{u_i^{*}}\right)=P\left(\lambda_i<U_i^{*}\le\lambda_{i+1}\right)\;،\;\lambda_q^{L_q+1}=+\infty\;\;،\;\lambda_q^{0}=-\infty
$$

$\boldsymbol{\omega}_{21}=\left(\boldsymbol{\Sigma}_{I_{\mathbf{u}_k},I_{u_{k1}^{*}}},\boldsymbol{\Sigma}_{I_{\mathbf{u}_k},I_{u_{k2}^{*}}},\boldsymbol{\Sigma}_{I_{\mathbf{u}_k},I_{n_k}}\right)$ و $\boldsymbol{\omega}_{11}$ به‌صورت زیر تعریف شده است:

$$
\boldsymbol{\omega}_{11}=\begin{pmatrix}
\boldsymbol{\Sigma}_{I_{u_{k1}^{*}},I_{u_{k1}^{*}}} & \boldsymbol{\Sigma}_{I_{u_{k1}^{*}},I_{u_{k2}^{*}}} & \boldsymbol{\Sigma}_{I_{u_{k1}^{*}},I_{n_k}}\\
 & \boldsymbol{\Sigma}_{I_{u_{k2}^{*}},I_{u_{k2}^{*}}} & \boldsymbol{\Sigma}_{I_{u_{k2}^{*}},I_{n_k}}\\
 & & \boldsymbol{\Sigma}_{I_{n_k},I_{n_k}}
\end{pmatrix}
$$

که در آن‌ها،

$$\boldsymbol{\Sigma}_{I_{u_i^*},I_{u_j^*}} = \operatorname{cov}\left(I_{u_i^*},I_{u_j^*}\right) = P\left(\lambda_i < U_i^* \le \lambda_{i+1}, \lambda_j < U_j^* \le \lambda_{j+1}\right) - P\left(\lambda_i < U_i^* \le \lambda_{i+1}\right) P\left(\lambda_j < U_j^* \le \lambda_{j+1}\right).$$

۲- ۱- ۳- رده‌بندی مدل مکانی عام جدید

در این بخش، ابتدا تابع چگالی توأم پاسخ‌های پیوسته، اسمی و ترتیبی را که از طریق مدل مکانی عام جدید به‌دست آمده است، مطرح نموده، سپس با استفاده از قاعده‌ی رده‌بندی ولچ [۲]، قاعده‌ی رده‌بندی برای آن ارائه می‌شود. برای این منظور، فرض کنیم $\mathbf{u}$، $\mathbf{n}$، $\mathbf{o}=(O_1,O_2)^T$ و $\mathbf{u}^*=\left(U_1^*,U_2^*\right)^T$ به ترتیب بردارهای متغیرهای پیوسته، اسمی، ترتیبی و پنهان باشند، با استفاده از مدل مکانی عام جدید، تابع چگالی توأم آن‌ها به‌صورت زیر بیان می‌شود $(d=1,2,...,D)$:

$$\begin{aligned} f_{\mathbf{O},\mathbf{U},\mathbf{N}}(\mathbf{o},\mathbf{u},\mathbf{n}) &= f(\mathbf{u}\mid\mathbf{o},\mathbf{n})\, f(\mathbf{o},\mathbf{n}) \\ &\cong \left\{ (2\boldsymbol{\pi})^{-0.5} \left|\boldsymbol{\Sigma}\right|^{-0.5} \exp\left\{ -\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_d)^T \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu}_d) \right\} \right. \qquad (۳) \\ &\quad \left. + \frac{\partial}{\partial \mathbf{u}} \boldsymbol{\omega}_{21} \boldsymbol{\omega}_{11}^{-1} \left(1-E\left(I_{\mathbf{u}_1^*}\right), 1-E\left(I_{\mathbf{u}_2^*}\right), 1-E\left(I_{\mathbf{n}}\right)\right)^T \right\} \boldsymbol{\pi}_d \end{aligned}$$

که در آن $\boldsymbol{\mu}_d$ بردار میانگین $\mathbf{u}$ برای خانه‌ی dام جدول پیشایندی و $\boldsymbol{\Sigma}$ ماتریس کوواریانس $\mathbf{u}$ است.

فرض کنیم $\boldsymbol{\Pi}^{(1)}$ و $\boldsymbol{\Pi}^{(2)}$ دو جامعه با احتمال‌های پیشین، هزینه‌های بد-رده‌بندی و ماتریس کواریانس برابر باشند. مشاهده‌ی آمیخته‌ی $\mathbf{w}'=\{\mathbf{o},\mathbf{u},\mathbf{n}\}$ طبق قاعده‌ی رده‌بندی ولچ [۲] و با استفاده از تابع چگالی توأم پاسخ‌های آمیخته‌ی به‌دست آمده در رابطه‌ی (۳) به جامعه‌ی $\boldsymbol{\Pi}^{(1)}$ اختصاص داده می‌شود هرگاه

$$\mathcal{R}_1: \quad \frac{\left[a\exp\left\{ -\frac{1}{2}\left(\mathbf{u}-\boldsymbol{\mu}_d^{(1)}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{u}-\boldsymbol{\mu}_d^{(1)}\right)\right\} + t \right]}{\left[a\exp\left\{ -\frac{1}{2}\left(\mathbf{u}-\boldsymbol{\mu}_d^{(2)}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{u}-\boldsymbol{\mu}_d^{(2)}\right)\right\} + t \right]} \ge \frac{\pi_d^{(2)}}{\pi_d^{(1)}} \qquad (۴)$$

و در غیر این صورت به جامعه‌ی $\boldsymbol{\Pi}^{(۲)}$ اختصاص داده می‌شود که در آن $a=(۲\pi)^{-۰/۵}\left|\boldsymbol{\Sigma}\right|^{-۰/۵}$ و $t=\frac{\partial}{\partial \mathbf{u}}\boldsymbol{\omega}_{۲۱}\boldsymbol{\omega}_{۱۱}^{-۱}\left(۱-E\left(I_{\mathbf{u}_۱^*}\right),۱-E\left(I_{\mathbf{u}_۲^*}\right),۱-E\left(I_{\mathbf{n}}\right)\right)^T$ مقادیر ثابتی هستند. $\boldsymbol{\mu}_d^{(g)}$ نیز میانگین متغیر پیوسته‌ی $\mathbf{u}$ در خانه‌ی dام جدول پیشایندی مربوط به جامعه‌ی gام $(g=۱,۲)$ است. به عبارت دیگر، $\mathbf{w}'$ به جامعه‌ی $\boldsymbol{\Pi}^{(۱)}$ اختصاص داده می‌شود هرگاه $\mathbf{w}'\in\mathcal{R}_۱$ و به جامعه‌ی $\boldsymbol{\Pi}^{(۲)}$ اختصاص داده می‌شود هرگاه $\mathbf{w}'\in\mathcal{R}_۲$. لازم به ذکر است که با استفاده از نمونه‌ی مستقل به اندازه‌ی $N^{(۱)}$ و $N^{(۲)}$ به ترتیب از جامعه‌های $\boldsymbol{\Pi}^{(۱)}$ و $\boldsymbol{\Pi}^{(۲)}$، با جایگذاری برآورد پارامترها (به دست آمده از تابع درستنمایی مربوطه) $\mathcal{R}_۱$ به‌صورت زیر برآورد می‌شود.

$$\hat{\mathcal{R}}_۱:\quad \frac{\left[(۲\pi)^{-۰/۵}\left|\mathbf{S}_p\right|^{-۰/۵}\exp\left\{-\frac{۱}{۲}\left(\mathbf{u}-\bar{\mathbf{u}}_d^{(۱)}\right)^T\mathbf{S}_p^{-۱}\left(\mathbf{u}-\bar{\mathbf{u}}_d^{(۱)}\right)\right\}+\hat{t}\right]}{\left[(۲\pi)^{-۰/۵}\left|\mathbf{S}_p\right|^{-۰/۵}\exp\left\{-\frac{۱}{۲}\left(\mathbf{u}-\bar{\mathbf{u}}_d^{(۲)}\right)^T\mathbf{S}_p^{-۱}\left(\mathbf{u}-\bar{\mathbf{u}}_d^{(۲)}\right)\right\}+\hat{t}\right]}\geq\frac{\hat{\pi}_d^{(۲)}}{\hat{\pi}_d^{(۱)}}\qquad (۵)$$

که در آن $\mathbf{S}_p$ ماتریس کواریانس نمونه‌ای ادغام‌شده، $\bar{\mathbf{u}}_d^{(g)}$ بردار میانگین نمونه‌ای متغیر پیوسته برای خانه‌ی d-ام در جامعه‌ی g-ام $(g=۱,۲)$ هستند. بنابراین، می‌توان چنین بیان کرد که مشاهده‌ی آمیخته‌ی $\mathbf{w}'=\{\mathbf{o},\mathbf{u},\mathbf{n}\}$ به جامعه‌ی $\boldsymbol{\Pi}^{(۱)}$ اختصاص داده می‌شود هرگاه رابطه‌ی (۵) برقرار باشد در غیر این صورت این مشاهده به جامعه‌ی $\boldsymbol{\Pi}^{(۲)}$ اختصاص داده خواهد شد. قاعده‌ی رده‌بندی بهینه‌ی مدل مکانی جدید را می‌توان برای موارد بیش از دو جامعه تعمیم داد. با فرض این‌که $\boldsymbol{\Pi}^{(۱)}$،...، $\boldsymbol{\Pi}^{(G)}$، $G\geq۲$، جوامعی باشند که فقط در بردار پارامترهای مکانی باهم اختلاف دارند، آنگاه طبق قاعده‌ی رده‌بندی بهینه‌ی ولچ [۲] و بر پایه‌ی مدل مکانی جدید، با فرض برابری هزینه‌های بد-رده‌بندی برای تمام جوامع، مشاهده‌ی آمیخته‌ی $\mathbf{w}'=\{\mathbf{o},\mathbf{u},\mathbf{n}\}$ به جامعه‌ای که برای آن

$$\begin{aligned}\delta^{(g)}&=\log\alpha^{(g)}+\log\pi_d^{(g)}\\&\quad+\log\left[a\exp\left\{-\frac{۱}{۲}\left(\mathbf{u}-\boldsymbol{\mu}_d^{(g)}\right)^{\top}\boldsymbol{\Sigma}^{-۱}\left(\mathbf{u}-\mu_d^{(g)}\right)\right\}+t\right],\ g=۱,۲,...,G\end{aligned}\qquad (۶)$$

بیش‌ترین مقدار را در بین سایر جوامع داشته باشد، اختصاص داده خواهد شد.

محاسبه‌ی احتمال بد-رده‌بندی. با فرض $G=۲$ و برابری احتمال‌های پیشین و هزینه‌ی بدرده‌بندی برای هر دو جامعه‌ی $\boldsymbol{\Pi}^{(۱)}$ و $\boldsymbol{\Pi}^{(۲)}$، احتمال بد-رده‌بندی بر پایه‌ی مدل مکانی عام جدید به‌صورت زیر بیان می‌شود:

$$\Pr(g \mid g') = \sum_{d}\sum_{\ell}\left(\int_{\hat{\mathcal{R}}_۲} I(g'=۱) + \int_{\hat{\mathcal{R}}_۱} I(g'=۲)\right)\left(\left\{\phi_C^{(g')}(\mathbf{u})+t\right\}\pi_d^{(g')}d\mathbf{u}\right) \quad (۷)$$

که در آن $\phi_C(\mathbf{u})$ تابع چگالی نرمال C- متغیره با بردار میانگین $\boldsymbol{\mu}_d$ و ماتریس کوواریانس $\boldsymbol{\Sigma}$ و $I(.)$ تابع نشان‌گر است.

محاسبه‌ی نرخ واقعی خطا. نرخ واقعی خطا (AER)[1] به‌صورت زیر بیان می‌شود:

$$AER = \frac{P(۱|۲)+P(۲|۱)}{۲}. \quad (۸)$$

برای برآورد میانگین رابطه‌ی (۸)، ابتدا برآورد ماکسیمم درستنمایی پارامترهای موجود در مدل مکانی عام جدید با استفاده از دستور "nlminb" در نرم‌افزار R که برای می‌نیمم کردن توابع غیرخطی و غیره استفاده می‌شود، به دست می‌آید و سپس انتگرال‌های موجود در (۷) با استفاده از روش‌های مونت‌کارلو تعیین می‌شوند. همچنین شیوه‌ی دیگری نیز به‌نام «برآورد جدا نگه داشته شده»[2] مطرح است، برای اطلاعات بیشتر به [۱۲] رجوع شود.

۳- مطالعه‌ی شبیه‌سازی

در این بخش، با استفاده از یک مطالعه‌ی شبیه‌سازی، قواعد رده‌بندی مدل مکانی عام دِ لئون و کریر [۱] و مدل مکانی عام جدید مورد بررسی قرارگرفته، نرخ خطای بد-رده‌بندی مدل مکانی عام دِ لئون و کریر [۱] را با نرخ خطای بد-رده‌بندی مدل مکانی عام جدید مورد مقایسه قرار داده‌ایم.

به منظور شبیه‌سازی، دو جامعه‌ی $\boldsymbol{\Pi}^{(۱)}$ و $\boldsymbol{\Pi}^{(۲)}$ با یک متغیر پیوسته، یک بردار اسمی با دو حالت و یک متغیر ترتیبی با دو سطح در نظر گرفته شده است. پارامترهای مدل نیز به‌

1- Actual Error Rate
2- Holdout Method

صورت $\boldsymbol{\Theta}_1=(\pi,\mu_1,\mu_2,\lambda)^T$ و $\boldsymbol{\Theta}_2=(\sigma^2,\rho,\beta)^T$ نشان داده‌شده‌اند که در آن μ_s بیانگر میانگین متغیر پیوسته‌ی U در خانه‌ی s، λ نقطه‌ی آستانه‌ای برای متغیر پنهان U^* تحت متغیر ترتیبی O است و همچنین ρ ضریب همبستگی بین U و U^* است. متغیر ترتیبی O به‌صورت زیر تعریف می‌شود:

$$O=\begin{cases}1 & U^*\le\lambda\\ 2 & U^*>\lambda\end{cases}$$

که در آن λ نقطه‌ی برشی برای متغیر پنهان U^* است. مشابه دِ لئون و کریر [۱] چهار حالت زیر در نظر گرفته‌شده و مورد بررسی قرار می‌گیرد:

(I) هر دو جامعه، در بردار اسمی با هم اختلاف دارند؛

(II) هر دو جامعه، در متغیر پیوسته باهم اختلاف دارند؛

(III) هر دو جامعه، در متغیر ترتیبی با هم اختلاف دارند؛

(IV) جوامع نسبت به همه متغیرها با هم اختلاف دارند.

برای اجرای قاعده‌ی رده‌بندی، نمونه‌های مستقل به اندازه‌های

$$(N^{(1)},N^{(2)})=(50,100),\ (100,100),\ (200,250)$$

برای دو جامعه‌ی $\boldsymbol{\Pi}^{(1)}$ و $\boldsymbol{\Pi}^{(2)}$ تولید می‌شود. برای تولید داده‌های مورد نظر، ابتدا متغیر اسمی X را از توزیع برنولی و سپس متغیرهای پیوسته و متغیر پنهان مربوط به متغیر ترتیبی را از توزیع نرمال دو متغیره با استفاده از مقادیر پارامتری زیر برای چهار حالت تولید می‌نماییم.

(I-IV) $g=1,2$ ، $\boldsymbol{\Theta}_2^{(g)}=(\sigma^2,\rho,\beta)^T=(1.0/5,0/58)^T$؛

(I) $\boldsymbol{\Theta}_1^{(1)}=(0/3,0,2,-1/15)^T$ و $\boldsymbol{\Theta}_1^{(2)}=(0/8,0,2,-1/15)^T$؛

(II) $\boldsymbol{\Theta}_1^{(1)}=(0/5,3,-1,2/31)^T$ و $\boldsymbol{\Theta}_1^{(2)}=(0/5,4,0,2/31)^T$؛

(III) $\boldsymbol{\Theta}_1^{(1)}=(0/5,0,0/5,-0/29)^T$ و $\boldsymbol{\Theta}_1^{(2)}=(0/5,0,0/5,-2/02)^T$؛

(IV) $\boldsymbol{\Theta}_1^{(1)}=(0/3,2/5,1/1,0/81)^T$ و $\boldsymbol{\Theta}_1^{(2)}=(0/8,0,-1/8,0/12)^T$.

ابتدا دو مدل مکانی عام دِ لئون و کریر [۱] و مدل مکانی عام جدید را روی داده‌های تولیدشده در چهار حالت برازش داده، برآورد پارامترهای هر یک از این مدل‌ها را به دست می‌آوریم. سپس تولید داده‌ها و برآورد پارامترهای هر دو مدل را ۱۰۰۰ بار تکرار می‌کنیم. در نهایت، متوسط تعداد نمونه‌های بد-رده‌بندی شده در هر دو جامعه را برای هر دو مدل با استفاده از احتمال بد-رده‌بندی شده به دست می‌آوریم. نتایج، در جدول (۱) آمده است.

جدول (۱) بیان‌گر متوسط تعداد افرادی است که در هر دو نمونه‌ی $N^{(۱)}$ و $N^{(۲)}$ بد-رده‌بندی شده‌اند. مثلاً برای نمونه‌ای به اندازه‌ی ۴۵۰ (جمعِ اندازه‌ی نمونه‌ی هر دو جامعه) در حالت (IV)، عدد ۲۳/۱۵ مربوط به مدل مکانی جدید در مقایسه با مدل دِ لئون و کریر [۱]، (عدد ۲۴/۳۳) نشان دهنده‌ی رده‌بندی بهتر (متوسط تعداد بد-رده‌بندی‌های پایین) است. این نتیجه برای چهار حالت بیان‌شده در نمونه‌های مختلف نیز صدق می‌کند و نشان دهنده‌ی این است که مدل مکانی عام جدید بهتر از مدل مکانی عام دِ لئون و کریر [۱] عمل می‌کند.

جدول ۱: متوسط تعداد نمونه‌های بد-رده‌بندی شده در هر دو جامعه

	$N^{(۲)}$	$N^{(۱)}$	مدل مکانی عام جدید	مدل مکانی عام دِ لئون و کریر [۱]
I	۱۰۰	۵۰	۳۸/۴۵	۴۰/۴۳
	۱۰۰	۱۰۰	۴۷/۵۶	۴۸/۵۹
	۲۵۰	۲۰۰	۱۱۳/۶۶	۱۱۵/۹۰
II	۱۰۰	۵۰	۴۱/۱۲	۴۳/۰
	۱۰۰	۱۰۰	۶۰/۸۷	۶۱/۴۴
	۲۵۰	۲۰۰	۱۳۳/۹۹	۱۳۴/۷۷
III	۱۰۰	۵۰	۳۵/۹۶	۳۷/۰۱
	۱۰۰	۱۰۰	۱۱/۱۲	۱۲/۵۲
	۲۵۰	۲۰۰	۲۵/۴۵	۲۷/۴۹
IV	۱۰۰	۵۰	۱۰/۱۴	۱۱/۲۱
	۱۰۰	۱۰۰	۱۱/۳۴	۱۲/۳۲
	۲۵۰	۲۰۰	۲۳/۱۵	۲۴/۳۳

دلیل اصلی آن، این است که در مدل عام جدید، ترتیبی بودن متغیر در مدل وارد شده و سبب ایجاد خانه‌های بیشتر در مدل مکانی عام شده است در صورتی که در مدل مکانی عام دِ لئون و کریر [۱]، ترتیبی بودن نقشی همانند متغیر پیوسته در جدول پیشایندی بازی می‌کند و نسبت به مدل مکانی عام جدید اطلاعات کم‌تری را در خود دارد و می‌توان

این‌گونه استنباط کرد که شیوه‌ی رده‌بندی با افزایش اطلاعات (در اینجا با اضافه کردن خانه‌های جدول پیشایندی) بهبود می‌یابد. در نتیجه در مسئله‌ی رده‌بندی مشاهدات، مدل مکانی عام جدید در مقایسه با مدل دِ لئون و کریر [۱]، توصیه می‌گردد.

۴- کاربرد روی داده‌های آمیخته‌ی بیماری کبد چرب

در این مقاله، از داده‌های بیماری کبد چرب که از مطالعه‌ای بر روی بیماران بیمارستان طالقانی تهران جمع‌آوری شده بود، استفاده شده است. بیماری کبد چرب، از تجمع زیاد چربی در اطراف کبد حادث می‌شود و سبب مختل شدن کار کبد در بدن می‌گردد. از سویی این بیماری در ایران بسیار شایع بوده و علائم خاصی نیز ندارد. به همین منظور با استفاده از متغیرهای موثر روی کبد چرب و با استفاده از مدل مکانی عام جدید به دنبال رده‌بندی درست بیماران مبتلا به کبد چرب هستیم. متغیرهای موثر درنظرگرفته‌شده در بیماری کبد چرب عبارتند از: میزان چاقی که بیان‌کننده نسبت وزن به مجذور قد افراد است (متغیر پیوسته)، وجود/عدم وجود علائم بیماری کبد چرب (متغیر اسمی، ۱: وجود علائم، ۲: عدم وجود علائم) و شدت بیماری کبد چرب (متغیر ترتیبی، با سه سطح ۱: فرد کبد چرب ندارد، ۲: فرد کبد چرب متوسطی دارد و ۳: فرد کبد چرب شدیدی دارد). بیش‌ترین تمرکز این مطالعه، بررسی و شناسایی در تفکیک بیماران نیازمند بستری و درمان سرپایی است. متغیرهای آمیخته براساس $N^{(1)} = ۱۴۶$ بیمار سرپایی (بیمارانی که ترخیص شدند، جامعه‌ی اول) و $N^{(2)} = ۳۸$ بیمار بستری (بیمارانی که در بیمارستان بستری شدند، جامعه‌ی دوم) مورد بررسی قرار گرفته است.

با استفاده از قاعده‌ی رده‌بندی حاصل از مدل مکانی عام جدید (رابطه‌ی ۴) و جایگزین کردن برآورد ماکسیمم درستنمایی پارامترهای مدل مکانی عام جدید، نتایج مربوط به قاعده‌ی رده‌بندی برای بیمار مراجعه‌کننده، به‌دست می‌آید. برای مثال، قاعده‌ی رده‌بندی برای بیمار مراجعه‌کننده به این صورت خواهد بود که اگر یک بیمار بدون شدت کبد چرب و با وجود علائم بیماری کبد چرب باشد، آنگاه برای بستری پذیرفته می‌شود هرگاه رابطه‌ی ۴ برقرار باشد در غیر این صورت بیمار ترخیص می‌شود. قواعد رده‌بندی برای همه‌ی وضعیت‌ها و سطوح متغیر کبد چرب و متغیر وجود/عدم وجود علائم بیماری کبد چرب برآورد می‌شوند. لازم به ذکر است که این نتایج برای استفاده در موقعیت‌های بالینی که نیاز به تصمیم‌گیری برای بستری یا ترخیص احساس می‌شود، می‌توانند به‌کاربرده شوند. نتایج رده‌بندی‌های صورت گرفته براساس مدل مکانی عام جدید و مدل مکانی دِ لئون و

کریر [۱]، با فرض برابری احتمال‌های پیشین و همچنین برابری هزینه‌های بد-رده‌بندی برای هر دو گروه بیمارها در جدول (۲) نشان داده شده است.

جدول ۲: تعداد (درصد) نمونه‌های بد-رده‌بندی شده مدل مکانی عام جدید و مدل دِلئون و کریر [۱] برای داده‌های کبد چرب

گروه	اندازه‌ی نمونه	مدل مکانی جدید	مدل مکانی عام دِ لئون و کریر[۱]
بیماران بستری‌شده	۳۸	۱۰ (۲۶/۳۱)	۱۳ (۳۴/۲۱)
بیماران سرپایی	۱۴۶	۴۵ (۳۰/۸۲)	۴۸ (۳۲/۸۷)
جمع	۱۸۴	۵۵ (۲۹/۸۹)	۶۱ (۳۳/۱۵)

نتایج جدول (۲) حاکی از آن است که در مدل مکانی دِ لئون و کریر [۱]، ۱۳ بیمار از گروه بیماران بستری‌شده در گروه بیماران سرپایی بد-رده‌بندی شده و به اشتباه ترخیص شده‌اند. با توجه به مقدار ۲۶/۳۱ درصد بیماران ترخیص شده که در نوع خود عدد بزرگی است برای بد-رده‌بندی بیمارانی که باید بستری شوند و نشده‌اند، باید به دنبال راهکارهای کاهش خطای بد-رده‌بندی با افزایش میزان بازدهی کارکنان و مسئولین بهداشتی و پزشکان ناظر باشیم. همچنین ملاحظه می‌شود در اینجا نیز مدل مکانی عام جدید، بهتر از مدل مکانی عام دِ لئون و کریر [۱]، مشاهدات آمیخته را رده‌بندی می‌نماید.

۵- بحث و نتیجه‌گیری

در این مقاله، با استفاده از مدل مکانی عام جدید توانستیم تفکیک و رده‌بندی داده‌های آمیخته را نسبت به مدل دِ لئون و کریر [۱]، بهبود بخشیم. رده‌بندی بهتر این مدل به خاطر افزایش خانه‌های جدول پیشایندی حاصل از ورود متغیر ترتیبی در ساختن جدول پیشایندی، ایجاد شده است. از سویی دیگر این موضوع تاکنون مورد بررسی قرار نگرفته است و می‌توان آن را به عنوان موضوع جدیدی در حوزه‌های زیر مورد توجه قرار داد: الف) برای پاسخ‌های دارای گم‌شدگی غیرقابل چشم‌پوشی؛ ب) برای پاسخ‌های طولی.

مراجع

[1] De Leon, A.R. and Carrière, K.C. (2007). General mixed-data model: Extension of general location and grouped continuous models. *The Canadian Journal of Statistics*, **35**, 533 - 548.

[2] Welch, B.L. (1939). Note on discriminant functions. *Biometrika*, **31**, 218–220.

[3] Tate, F.R. (1954). Correlation between discrete and continuous variables: Point-Biserial correlation. *Ann. Math. Statist.*, **25**, 603-607.

[4] Cox, D.R. and Wermuth, N. (1992). Response models for mixed binary and quantative variables. *Biometrika*, **79**(3), 441-461.

[5] Catalano, P. and Ryan, L.M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association,* **87**, 651-658.

[6] De Leon, A.R., Soo, A. and Williamson, T. (2011). Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistics*, **5** (38), 1021-1032.

[7] Trivedi, P. and Zimmer, D. (2006). *Copula modelling in econometrics: Introduction to practitioners*. Wiley, New York.

[8] Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G. and Bijnes, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Stat. Med*., **27**, 4408-4427.

[9] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag. New York.

[10] Sammel, M.D. Ryan, L.M. and Legler J.M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society,* **B**, 59, 667–678.

[11] Joe, H. (1995). Approximations multivariate normal rectangle probabilites based on conditional expectation. *Journal of the American Statistical Association*, **90**: 957-967.

[12] Johnson, R.A. and Wichern, D.W. (2007). *Applied multivariate statistical analysis*. 6th ed., Prentice Hall. New York.

Analysis of Mixed Discrete and Continuous Classified Responses

Ehsan Bahrami Samani and Hamid Reza Aalipour

Department of Statistics, Shahid Beheshti University, Tehran, Iran

Abstract

In this paper, classification of mixed discrete and continuous responses is our goal. To do this, first we have the joint distribution function of these data. Therefore, we propose new general location model to obtain the joint distribution function of mixed discrete and continuous responses and briefly compare it with the De Leon and Carrière's general location model [1]. The classification approach is based on Welch's classification rules [2] by providing misclassification probabilities and actual error rate. In addition, the results of classification of the new general location model are compared with the same one of the De Leon and Carrière's general location model [1]. It is noted that, the sample classification rules of new general location model are provided through obtaining parameters estimation of the maximum likelihood function. Finally, to show the performance of the rules, simulation study and analysis of a real data set using new general location model, is performed.

Keywords: Classification rule; General location model; Continuous and Discrete mixed data; Likelihood Function; Misclassification Probability.

Mathematics Subject Classification (2010): 62J12, 62J05.