تعیین ماتریس طرح در مدل‌های نسبتی به کمک مقابله‌های تحت فرض در جداول پیشایندی

سید کامران قریشی

گروه آمار دانشگاه قم

چکیده: مدل‌هایی که اخیراً برای تحلیل ساختار جدول‌های پیشایندی اعم از جداول ناقص یا کامل به کار می روند مدل‌های نسبتی نامیده می‌شوند. اساس کار این مدل‌ها بر پایه ماتریس طرحی است که از پیش تعیین شده باشند. در عمل ممکن است به جای داشتن این ماتریس فرضهایی در خصوص نسبت بختهای تعمیم یافته وجود داشته باشد. این مقاله به مساله یافتن ماتریس طرح بر اساس فرضیات داده شده خواهد پرداخت.

واژه‌های کلیدی: جدول‌های پیشایندی، مدل‌های ضربی، مدل‌های لگ خطی، مدل‌های نسبتی

۱ مقدمه

مدل‌های لگ خطی به عنوان شق دیگری از مدل‌های ضربی در تحلیل جداول پیشایندی به کار می‌روند. در مدل‌های لگ خطی معمولاً مقادیر مختلف متغیرهای سازنده جدول نقشی اساسی در تعیین الگوی آن دارند. اما در عمل ممکن است بعضی از خانه‌های جدول از ویژگی خاصی نیز برخوردار باشند که در این صورت لازم است این ویژگی در ساختار مدل به کار رود. علاوه بر آن به کارگیری مدل‌های لگ خطی در جدول‌های پیشایندی با صفر ساختاری با پیچیدگی‌های خاصی روبروست. هر چند با استفاده از مدل‌های زیر فضای سلسله مراتبی HSM [۱] که بر پایه گراف‌ها پایه‌ریزی شده‌اند تا حد بالایی می‌توان بر این مشکل فایق آمد،((Hisayuki, et al.(2012).
اخیراً در مقاله‌ای مدل‌هایی برای جداول پیشایندی معرفی شده است که مدل‌های نسبتی RM [۲] نامیده می‌شوند،((Klimova, et al.(2012). این مدل‌ها به عنوان حالت خاص شامل مدل‌های لگ خطی ((Agresti, (2002) و مدل‌های ضربی ((Goodman, (1972) هستند و علاوه بر آن برای تحلیل هر جدول پیشایندی اعم از ناقص یا کامل کاربرد

[۱] Hierarchial Subspace Models
[۲] Relational Models

دارند. در این مقاله، که تنها مقاله در این زمینه است، خواص کلی مدل‌های نسبتی بحث شده است. به عنوان نمونه این که تحت چه شرایطی یک مدل نسبتی یک مدل نمایی معمولی و تحت چه شرایطی مدل نمایی خمیده است به طور مفصل بسط یافته و وجود و یکتایی برآوردهای ماکسیمم درستنمایی mle نیز به تفصیل بررسی شده است. اما مسائل اساسی نظیر این که برآوردهای ماکسیمم درستنمایی تحت چه شرایطی دارای نمایش مکانی هستند و اصولاً تحت فرض داده شده H_0، چگونه می‌توان ماتریس طرح را مشخص کرد موارد قابل بحث است. به دلیل اهمیت تعیین ماتریس طرح، ما در این مقاله به چگونگی تعیین آن، بر اساس مقابله‌های داده شده، خواهیم پرداخت.
در بخش دوم این مقاله به طور مختصر با مدل‌های نسبتی آشنا می‌شویم. چگونگی تعیین ماتریس طرح در بخش سوم به تفکیک برای مقابله‌های نوع اول و دوم بحث خواهد شد. در پایان نتایج حاصل از بخش ۳ را برای دو جدول پیشایندی تحت عنوان دو مثال به کار خواهیم بست.

۲ نمادگذاری و تعاریف

فرض کنید $X_1,...,X_k$ متغیرهای رسته‌ای با فضای سطوح $\chi_1,...,\chi_k$ باشند که‌به‌ترتیب در $I_1,...,I_k$ سطح واقع‌اند. در این جا فرض می‌کنیم، $I_i \geq 2; i = 1,...,k$. نقطه $(x_1,...,x_k) \in \chi_1 \times ... \times \chi_k$ یک خانه تولید می‌کند اگر و فقط اگر پیشامد $(x_1,...,x_k)$ در جامعه رخ دهد. یک خانه $(x_1,...,x_k)$ خالی است اگر در طرح نباشد. معمولاً چنین خانه‌هایی را خانه‌های با صفر ساختاری گویند. بدون کاستن از کلیت مساله، فرض کنید $I = I_1 \times I_2, ... \times I_k$ و $|I|$ نمایش تعداد خانه‌های غیر خالی باشد. همچنین فرض کنیم خانه‌های جدول بر اساس سطوح متغیرهای رسته‌ای به صورت فرهنگ لغت نامه‌ای مرتب شده باشند. بر پایه اینکه داده‌ها بر اساس چه طرح نمونه‌گیری به دست آمده باشند پارامتر خانهiام را با $\delta(i)$ نمایش می‌دهیم که برای توزیع چند جمله‌ای $\delta(i) = p(i)$ و برای توزیع پواسون برابراست با $\delta(i) = \lambda(i)$. همچنین $\mathbf{S} = \{S_1,...,S_J\}$ را کلاسی از زیر مجموعه‌های غیر تهی از خانه‌های جدول در نظر می‌گیریم. لازم به ذکر است که هرگاه با یک جدول پیشایندی معمولی روبرو باشیم S_j، سیلندرها، یعنی همان خانه‌ها با جمع‌های حاشیه‌ای متناظر با آماره‌های بسنده خواهند بود. مثلا در یک جدول پیشایندی 2×2 داریم $S_1 = \{(1,1),(1,2)\}$ و $S_2 = \{(1,1),(2,1)\}$.
درایه‌های ماتریس طرح A، یعنی a_{ji} را مساوی یک تعریف می‌کنیم اگر خانه iام در سیلندر j باشد و در غیر این صورت برابر صفر در نظر می‌گیریم. با این مقدمه مدل

نسبتی عبارتند از:

$$log\delta = A'\beta, \qquad (۱)$$

که در آن $\beta \in R^{|J|}$.
اگر فرض کنیم که زیر مجموعه S_j برای هر$(j = ۱,...,|I|)$ تنها شامل یک عضو(یا یک خانه از جدول) باشد در این صورت ماتریس A برابر ماتریس واحد $|I| \times |I|$ بوده و مدل نسبتی را اشباع شده گوییم. در صورتی که ماتریس A پر رتبه سطری نباشد فضای متعامد آن را با نماد $D = ker(A)$ تعریف می‌کنیم. با این تعریف، ماتریس $M = \begin{pmatrix} A \\ D \end{pmatrix}$ معکوس پذیر بوده و داریم: $DA' = AD' = ۰$.
با توجه به تعریف D، نمایش معادل مدل نسبتی (۱) برابر است با:

$$Dlog\delta = ۰. \qquad (۲)$$

مدل (۲) در واقع از ضرب طرفین مدل (۱) در D به دست می آید به طوری که آن را نمایش لگاریتم نسبت بخت‌های تعمیم یافته نیز می‌نامیم. مدل‌های (۱) و (۲) لازم و ملزوم یکدیگرند به این معنی که داشتن یکی از آن‌ها دیگری را نتیجه می‌دهد.
مثال ۱. جدول پیشایندی یک بعدی با چهار خانه را در نظر بگیرید. مقادیر احتمال خانه‌ها با بردار $p = (p_۱, p_۲, p_۳, p_۴)$ داده می‌شود. فرض کنید بخواهیم درستی فرض $H_۰ : \frac{p_۲}{p_۳} = ۱$ یا به طور معادل $H_۰ : \log p_۲ = \log p_۳$ را بیازماییم. ماتریسهای A و D متناظر عبارت خواهند بود از:

$$A = \begin{pmatrix} ۱ & ۰ & ۰ & ۰ \\ ۰ & ۱ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۱ \end{pmatrix}, \; D = (۰, ۱, -۱, ۰).$$

کیلیموا و همکاران(۲۰۱۲) نشان دادند که در صورت وجود ماتریس A، و اگر بردار یکه درفضای‌سطری این ماتریس باشند، برآوردهای ماکسیم درستنمایی در رابطه $A\mathbf{n} = nAp$ صدق می‌کنند که در آن $\mathbf{n}$ بردار بسامدهای مشاهده شده، n اندازه نمونه و p بردار احتمال خانه‌های جدول می‌باشد که به صورت فرهنگ لغت نامه‌ای مرتب شده‌اند. لازم به ذکر است که در این حالت بسامدهای مورد انتظار خانه‌های جدول پیشایندی، تحت الگوهای نمونه‌گیری چندجمله‌ای و پواسون، برابرخواهندبود، ((Agresti, (2002).
اکنون با این مقدمه نسبتاً فشرده و آشنایی کلی با مدلهای نسبتی، آماده‌ایم تا به موضوع اصلی این مقاله یعنی تعیین ماتریس طرح بپردازیم.

۳ تعیین ماتریس طرح

همانطور که در بخش دوم دیدیم اساس مدل‌های نسبتی بر پایه ماتریس طرح یعنی ماتریس A بنا شده است. در صورتی که این ماتریس از قبل مشخص باشد تحلیل داده‌های رسته‌ای تقریباً سرراست خواهد بود. در اینجا از کلمه تقریباً استفاده شد زیرا اگر بردار یکه در فضای سطری ماتریس طرح باشد برآوردهای ماکسیمم درستنمایی برابر با مقادیر امید ریاضی آن‌ها خواهند بود در حالی که اگر بردار یکه در فضای سطری ماتریس طرح نباشد برآوردهای ماکسیمم درستنمایی در معادلاتی صدق می‌کنند که در آن‌ها مقادیر آماره‌های بسنده متناسب با مقادیر امید ریاضی آن‌ها خواهند بود. در این حالت اگر چه یافتن برآوردهای ماکسیمم درستنمایی برای بعضی از مسائل خاص به راحتی قابل انجام است اما در حالت کلی با پیچیدگی محاسباتی روبرو خواهیم بود.
باید توجه داشت که به غیر از موارد خاص (که شامل انواع استقلال در جدول‌های پیشایندی است) ساختار کلی ماتریس A با توجه به فرضیات موجود بین احتمال خانه های جدول تعیین می‌شود. این فرضیات گاهی ممکن است در خصوص جداول یک طرفه یا جداول پیشایندی ناقص باشند. اما این که چگونه می‌توان درایه‌های ماتریس A را با فرض‌های داده شده به دست آورد مساله‌ای است که در ادامه به آن می‌پردازیم.
تعریف ۱. اعمال قیود بر لگاریتم احتمال خانه‌های جدول ممکن است در قالب ترکیباتی به صورت

$$H_0 : \Sigma_{i=1}^{|I|} d_i(j) log p_i = 0;\ j = 1, ..., |I| - J, \qquad (۳)$$

باشند که در آن‌ها ضرایب $d_i(j)$ کمیت‌های معلوم فرض می‌شوند. اگر این ضرایب در شرط‌های $\Sigma_{i=1}^{|I|} d_i(j) = 0;\ j = 1, ..., |I| - J,$ صدق کنند، (۳) یک مقابله یا یک نسبت بخت تعمیم یافته نامیده می‌شود.
تعریف می‌کنیم:

$$\mathbf{d}(j) = (d_1(j), ..., d_{|I|}(j));\ j = 1, ..., |I| - J.$$

باید توجه داشت که تعداد مقابله‌های تحت فرض H_0 برابر تعداد درجات آزادی آماره خی دو است. اگر برای j خاص فرض کنیم تمام مؤلفه‌های بردار $\mathbf{d}(j)$، به غیر از دو مؤلفه که یکی برابر ۱ و دیگری برابر ۱- است، برابر صفر باشند، به آن بردار مقابله ساده نوع اول گوییم. مقابله ساده نوع دوم وقتی به دست می‌آید که به غیر از چهار مؤلفه، دوتای آن‌ها برابر ۱ و دوتای دیگر برابر ۱-، همگی برابر صفر باشند. مقابله‌هایی که از نوع اول و دوم نباشند مقابله‌های نوع سوم نامیده می‌شوند. به راحتی می‌توان ثابت کرد که برای مقابله‌های نوع اول و دوم، بردار یکه در فضای سطری ماتریس

طرح است و معادلات برآوردهای درستنمایی مدل مورد بررسی از شکل کلی مدل‌های نمایی معمولی تبعیت می‌کنند اما چنین تضمینی برای مقابله‌های نوع سوم وجود ندارد. لذا ما در این مقاله تنها به بررسی مقابله‌های نوع اول و دوم خواهیم پرداخت.

فرض کنیم $\mathbf{d}(j)$ ضرایب مقابله نوع اول یا دوم باشند. بدون کاستن از کلیت مساله برای مقابله نوع اول بگیریم: $\mathbf{d}(j) = (۱, -۱, ۰, ..., ۰)$. همچنین برای مقابله نوع دوم قرار می‌دهیم: $\mathbf{d}(j) = (۱, -۱, -۱, ۱, ۰, ..., ۰)$.

تعریف ۲. دو مقابله $\Sigma_{i=۱}^{|I|} d_i(j)\log p_i = ۰$ و $\Sigma_{i=۱}^{|I|} d_i(j')\log p_i = ۰$ را مقابله های مستقل تعریف می‌کنیم هر گاه بردار ضرایب آن‌ها عمود باشند. یعنی داشته باشیم: $\Sigma_{i=۱}^{|I|} d_i(j)d_i(j') = ۰;\ j \neq j'$.

اگر تحت فرض $H_۰$ مقابله‌ها مستقل باشند بلافاصله می‌توان نتیجه گرفت که در جدول پیشایندی، خانه‌های سازنده هر مقابله جدا از خانه‌های سازنده مقابله دیگر است. در صورت مستقل نبودن مقابله‌ها، ممکن است بعضی خانه‌ها در دو مقابله مشترک باشند. به عنوان مثال در یک جدول پیشایندی ۳ × ۳ دو مقابله نوع دوم $\log p_{۱۲} - \log p_{۲۲} + \log p_{۲۳} = ۰$ و $\log p_{۲۱} - \log p_{۳۱} - \log p_{۲۲} + \log p_{۳۲} = ۰$ را در نظر بگیرید. به دلیل اینکه خانه (۲, ۲) در هر دو مقابله شرکت دارد در نتیجه دو مقابله مستقل نیستند. از آن جا که هدف این مقاله تعیین ماتریس طرح برای مقابله‌های مستقل و غیر مستقل، و همچنین مقابله‌های نوع اول و دوم است، که شامل چهار حالت مختلف است، برای انسجام بیشتر و شفافیت در هر مورد بحث را در دو زیر بخش به تفکیک مقابله نوع اول و دوم ارائه می‌کنیم:

۱.۳ مقابله‌های نوع اول

دو مقابله نوع اول را در نظر بگیرید. ممکن است این دو مقابله مستقل بوده یا حداکثر در یک نقطه مشترک باشند. با بررسی این دو حالت نتایج برای بیش از دو مقابله نیز قابل تعمیم است.

برای ساختن ماتریس طرح A از الگوریتم زیر استفاده می‌کنیم:

۱) ابتدا یک ماتریس واحد $U = |I| \times |I|$ را در نظر بگیرید. سطرها و ستونهای این ماتریس را به ترتیب با شماره خانه جدول شماره‌گذاری می‌کنیم.

۲)

(a) اگر دو مقابله مستقل باشند برای هر مقابله به طور جداگانه دو سطر از ماتریس U، که شماره آن‌ها با اندیس جملات غیر صفر مقابله برابر است، را جمع می‌کنیم. به این ترتیب هر مقابله مستقل یک سطر از سطرهای ماتریس U را

کاهش می‌دهد.

(b) در صورتی که دو مقابله مستقل نباشند، احتمال‌های سه خانه از جدول پیشایندی، مقابله‌ها را می‌سازند پس در این صورت سه سطر از ماتریس U متناظر با شماره خانه‌های مذکور را با یکدیگر جمع می‌کنیم. دقت داریم که در این صورت دو سطر از ماتریس U حذف می‌شود. تعداد سطرهای حذف شده برابر تعداد درجات آزادی آماره خی دو است.

اگر مراحل ۱ و ۲ را برای همه مقابله‌های تحت فرض H_0 انجام دهیم ماتریس طرح تحت آن فرض حاصل می‌شود که از بعد J بوده و همواره بردار یکه در فضای سطری آن خواهد بود.

۲.۳ مقابله‌های نوع دوم

دو مقابله خاص نوع دوم را در نظر بگیرید. ممکن است این دو مقابله مستقل بوده یا در یک یا حداکثر دو نقطه مشترک باشند. با بررسی این حالت‌ها نتایج برای بیش از دو مقابله نیز قابل تعمیم است.

مجدداً الگوریتم زیر را برای ساختن ماتریس طرح A بکار می‌بریم:

۱) ابتدا یک ماتریس واحد $U = |I| \times |I|$ را در نظر می‌گیریم. سطرها و ستونهای این ماتریس را به ترتیب با شماره خانه جدول شماره‌گذاری می‌کنیم.

۲)

(a) اگر دو مقابله مستقل باشند برای هر مقابله به طور جداگانه چهار سطر از ماتریس U، که شماره آن‌ها با اندیس جملات غیر صفر مقابله برابر است، را انتخاب می‌کنیم. از روی این چهار سطر سه سطر جای گزین را به این ترتیب می‌سازیم: اولین سطر از جمع همه سطرهای ماتریس، دومین سطر از جمع سطرهای اول و دوم و در نهایت سومین سطر از جمع سطرهای اول و سوم ماتریس U حاصل می‌شوند. این عمل باعث می‌شود که برای هر مقابله تنها یک سطر از سطرهای ماتریس U کاسته شود.

(b) در صورتی که دو مقابله در یک نقطه(خانه) مشترک باشند در این صورت احتمال‌های هفت خانه‌ی جدول در این دو مقابله شرکت دارند. بنابر این باید دو سطر از هفت سطر ماتریس U به خاطر این دو مقابله کاسته شود که سازگار با دو درجه آزادی آماره خی دو است. بدون کاستن از کلیت مساله می‌توان ضریب جمله مشترک در هر دو مقابله را ۱+ گرفت. هفت سطر متناظر با شماره خانه‌های موجود در مقابله‌ها را در نظر می‌گیریم از روی این سطرها پنج بردار

جای گزین به صورت زیر می‌سازیم:
الف) بردار اول: از جمع سطرهای مذکور به دست می‌آید.
ب) بردار دوم: از جمع سطری که شماره آن با شماره خانه مشترک است با دو سطر با شماره‌های ۱- از هر یک از مقابله‌ها.
ج) بردار سوم: از جمع سطری که شماره آن با شماره خانه مشترک است با دو سطر با شماره‌های ۱- از هر یک از مقابله‌ها که در قسمت (ب) از آن‌ها استفاده نشده است.
د) بردار چهارم: از جمع دو سطر که شماره آن ها در مقابله اول بوده و ضریب آن‌ها مخالف و هیچ کدام از خانه مشترک نباشند.
ه) بردار پنجم: از جمع دو سطر که شماره آن ها در مقابله دوم بوده و ضریب آن‌ها مخالف و هیچ کدام از خانه مشترک نباشند.

(c) در صورتی که دو مقابله در دو نقطه مشترک باشند تعداد خانه‌هایی که مقابله‌ها را می‌سازند شش خانه است. به ازای دو مقابله مورد نظر دو درجه آزادی از مدل اشباع شده کسر می‌شود. بنابراین لازم است به جای شش سطر ماتریس U، که متناظر با خانه‌هایی هستند که در ساختن مقابله‌ها شرکت دارند، از چهار سطر جای گزین استفاده کرد. سطرهای مذکور به صورت زیر تعیین می‌شوند:

i. خانه‌های مشترک در هر دو مقابله مختلف العلامة هستند. در این صورت بدون کاستن از کلیت مساله فرض می‌کنیم خانه دوم و پنجم مشترکند و مقابله‌ها، با لگاریتم نسبت بخت‌های مجاور، جدول ۳×۲ زیر را می‌سازند:

$$\begin{pmatrix} p_1 & p_2 & p_3 \\ p_4 & p_5 & p_6 \end{pmatrix},$$

در این صورت چهار بردار عبارتند از:
الف) بردار اول: تمام ستون‌های متناظر با احتمال‌های خانه‌های ۶-۱ را مساوی ۱ و بقیه ستون‌ها را مساوی صفر قرار می‌دهیم.
ب) بردار دوم: خانه‌هایی که در ردیف اول ماتریس فوق هستند مساوی ۱ و بقیه ستون‌ها را مساوی صفر قرار می‌دهیم.
ج) بردار سوم: خانه‌هایی که در ستون اول ماتریس فوق هستند مساوی ۱ و بقیه ستون‌ها را مساوی صفر قرار می‌دهیم.
د)بردار چهارم: خانه‌هایی که در ردیف دوم ماتریس فوق هستند مساوی ۱ و بقیه ستون‌ها را مساوی صفر قرار می‌دهیم.

ii. در صورتی که خانه‌های مشترک در هر دو مقابله متحد العلامة باشند نمی‌توان عناصر ماتریس A را تشکیل داد زیرا در این صورت ماتریس D بر ماتریس A عمود نخواهد بود.

به عنوان کاربرد الگوریتم‌های ارائه شده در زیر بخشهای ۱.۳ و ۲.۳ دو مثال زیر را در نظر بگیرید:

مثال ۲.(رجوع به مثال ۱) بر اساس فرضیات مثال ۱، می‌خواهیم درستی فرض $H_۰ : p_۱ = p_۲ = p_۳$ یا به طور معادل، درستی فرض $H_۰ : \log p_۱ = \log p_۲,\ \log p_۱ = \log p_۳$ را بیازماییم. مقابله‌های این فرض از نوع اول و غیر مستقل‌اند. پس ماتریس طرح A و ماتریس متعامد D عبارتند از:

$$A = \begin{pmatrix} ۱ & ۱ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۱ \end{pmatrix},\ D = \begin{pmatrix} ۱ & -۱ & ۰ & ۰ \\ ۱ & ۰ & -۱ & ۰ \end{pmatrix}.$$

مثال ۳. جدول ناقص ۳ × ۳ زیر را در نظر بگیرید:

$$A = \begin{pmatrix} - & p_۱ & p_۲ \\ p_۳ & p_۴ & p_۵ \\ p_۶ & p_۷ & - \end{pmatrix}.$$

فرض کنید می‌خواهیم درستی فرض $H_۰ : \frac{p_۱p_۵}{p_۲p_۴} = ۱,\ \frac{p_۳p_۷}{p_۶p_۴} = ۱$ یا به طور معادل فرض

$$H_۰ : \log p_۱ + \log p_۵ - \log p_۲ - \log p_۴ = ۰,\ \log p_۳ + \log p_۷ - \log p_۶ - \log p_۴ = ۰$$

را بیازماییم. این دو مقابله از نوع دوم بوده و خانه (۲,۲) جدول در هر دو مقابله مشترک است لذا ماتریس طرح A و ماتریس متعامد D به صورت زیر خواهند بود:

$$A = \begin{pmatrix} ۱ & ۱ & ۱ & ۱ & ۱ & ۱ & ۱ \\ ۱ & ۱ & ۰ & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۱ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۰ & ۰ & ۱ & ۰ \\ ۱ & ۰ & ۰ & ۱ & ۰ & ۰ & ۱ \end{pmatrix},$$

و

$$D = \begin{pmatrix} ۱ & -۱ & ۰ & -۱ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۱ & -۱ & ۰ & -۱ & ۱ \end{pmatrix}.$$

مراجع

Agresti, A. (2002), *Categorical data analysis*, Wiley, Now York.

Hara, H. and Sei, T and Takemura, A. (2012), Hierarchical subspace models for contingency tables, *J. of Multivariate analysis*, **103**, 19-34.

Goodman, L.A. (1972), some multiplicative models for the analysis of cross-classified data in: *proceedings of the sixth Berkely Symposium of mathematical statistics and probability*

Kimova, A. and Rudas, T. and Dobra, A. (2012), Relational models for contingency tables, *J. of Multivariate analysis*, **104**, 159-173.