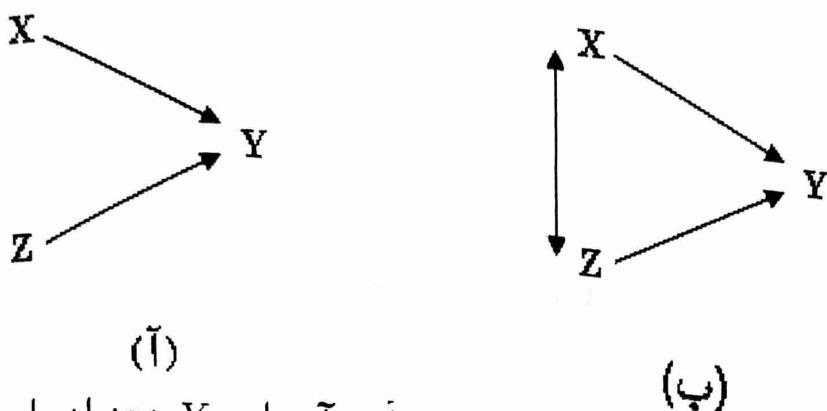


## فصل ۵

# مدل‌های لگ‌خطی برای جدول‌های پیش‌ایندی

مدل‌های لگ‌خطی برای مدل‌بندی شمارش‌های خانه‌ها در جدول‌های پیش‌ایندی، بسیار مورد استفاده‌اند. در این مدل‌ها، تأثیر متغیرهای رده‌بندی‌شده برآمید ریاضی تعداد شمارش‌های خانه‌ها مورد کاوش قرار می‌گیرد. در این مدل‌ها، هم اثرهای اصلی بین متغیرها برآمید ریاضی تعداد شمارش‌ها ارزیابی می‌شوند و هم اثرهای متقابل. توجه داشته باشید که در فصل‌های قبلی، متغیر پاسخ، یکی از متغیرهای مورد مطالعه (متغیرهای دودویی، ترتیبی یا اسمی) بود و تأثیر متغیرهای دیگر به عنوان متغیرهای تبیینی بر این متغیر، مورد بررسی قرار می‌گرفت. در این فصل، تعداد شمارش‌ها به عنوان یک متغیر پاسخ شمارشی که دارای توزیع پواسون یا چندجمله‌ای است، در نظر گرفته می‌شود و تأثیر متغیرهای دیگر به عنوان متغیرهایی که درون‌زا هستند (به این معنا که هر یک می‌توانند بر دیگری تأثیرگذار باشند) بر این پاسخ مورد علاقه، مورد تحلیل آماری قرار می‌گیرد. شکل ۱.۵ برای سه متغیر  $X$  و  $Y$  و  $Z$  تفاوت فصل‌های قبلی و این فصل را به نمایش می‌گذارد. در این فصل، نخست در بخش بعد، توزیع پواسون به عنوان توزیع تعداد شمارش‌ها در خانه‌ها بازنگری می‌شود. توزیع چندجمله‌ای و ارتباط آن با توزیع پواسون



شکل ۱.۵: (آ)  $X$  و  $Z$  دو متغیر تبیینی‌اند که تأثیر آن‌ها بر  $Y$  به عنوان پاسخ، مدل نظر است، (ب) در یک مدل لگ خطی، هر یک از متغیرها می‌تواند بر دیگری تأثیرگذار باشد

نیز در این بخش، مورد بررسی قرار می‌گیرد. سپس مدل‌های لگ خطی به عنوان حالت خاصی از مدل‌های خطی تعمیم‌یافته در بخش دوم مورد بحث قرار می‌گیرند. در این بخش، در خصوص مدل لگ خطی برای یک جدول پیش‌ایندی دو طرفه نیز بحث می‌شود. مدل لگ خطی برای جدول‌های چند طرفه در بخش‌های سوم و چهارم آورده شده است. در بخش پنجم، آزمون نیکویی برازش مدل بحث شده است. در بخش ششم، روش‌های برآورد پارامترها در مدل‌های لگ خطی مورد بحث و بررسی قرار می‌گیرند. در بخش هفتم، یک مثال کاربردی ارائه شده است. در انتهای در بخش هشتم، دستورهای R برای برازش مدل‌های لگ خطی آورده شده است.

## ۱.۵ توزیع چند جمله‌ای و ارتباط آن با توزیع پواسون

توزیع چند جمله‌ای، تعمیمی از توزیع دو جمله‌ای است به گونه‌ای که جامعه را به جای دو رده به  $k$  رده تقسیم می‌کنیم. به عبارت دیگر اگر  $N$  فرد به طور مستقل به  $2 > k$  رده قابل تفکیک باشند، تعداد افراد در رده‌ها دارای توزیع چند جمله‌ای است. در این توزیع اگر  $p_i$  و  $n_i$  برای  $i = 1, \dots, k$  به ترتیب، احتمال انتخاب و تعداد مشاهدات در رده‌ی  $i$ ام باشند، داریم  $\sum_{i=1}^k p_i = 1$  و  $\sum_{i=1}^k n_i = N$ . بنا بر این اگر

$n = (n_1, n_2, \dots, n_k) \sim \text{Multi}(N, p_1, \dots, p_k)$  باشد، خواهیم داشت:

$$P[n|N, p] = \frac{N!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k p_i^{n_i},$$

که در آن  $p = (p_1, \dots, p_k)$  دامنه‌ی مقادیر  $n = (n_1, n_2, \dots, n_k)$  عبارت است از  $\{n = (n_1, n_2, \dots, n_k) | \forall i = 1, \dots, k; n_i \in \{0, \dots, N\}, \sum_{i=1}^k n_i = N\}$ . همچنین در این توزیع،

$$E(n_i) = Np_i, \quad \text{var}(n_i) = Np_i(1 - p_i), \quad \text{cov}(n_i, n_j) = -Np_i p_j.$$

علت منفی بودن کوواریانس، این است که چون مجموع  $n_i$ ‌ها برابر با  $N$  است، با افزایش تعداد افراد در یک رده، تعداد افراد در دیگر رده‌ها کاهش می‌یابد. نکته‌ی دیگر این که زمانی که  $N$  بزرگ است، توزیع چندجمله‌ای با توزیع نرمال چندمتغیره تقریب زده می‌شود و داریم:

$$\frac{n-Np}{\sqrt{N}} \sim \text{MVN}(0, \Sigma),$$

که در آن، درایه‌های روی قطر  $\Sigma$  برابر با  $(1 - p_i)p_j$  و درایه‌های خارج قطر آن  $-p_i p_j$  هستند. بنا بر این  $\frac{n-Np}{\sqrt{N}}$  دارای توزیع نرمال چندمتغیره با میانگین صفر است. واریانس هر درایه  $(1 - p_i)p_j$  و کوواریانس مؤلفه‌های  $q_i$  و  $z_j$ ،  $-p_i p_j$  است.

گاهی داده‌های شمارشی به تعداد از قبل معینی منجر نمی‌شود. به عنوان مثال اگر  $y$  تعداد مرگ‌ها به واسطه‌ی تصادفات اتومبیل در هفته‌ی آینده باشد، حد بالای ثابت  $n$  برای  $y$  وجود ندارد. چون  $y$  می‌تواند مقادیر یک مجموعه‌ی شمارش‌پذیر را اخذ کند، ساده‌ترین توزیع برای این نوع داده‌ها، توزیع پواسون است که احتمال‌ها به تک‌پارامتر میانگین  $\mu$  وابسته است. تابع جرم احتمال پواسون (پواسون، ۱۸۳۷، ص. ۶۰) به صورت زیر تعریف می‌شود:

$$\Pr(Y = y) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

که از آن داریم  $E(Y) = \mu = \text{var}(Y)$ . توزیع پواسون، یک توزیع تک‌مُدی با مُدی برابر با عدد صحیح  $\mu$  است. چولگی در این توزیع،  $\frac{E(Y-\mu)^2}{\sigma^2} = \frac{1}{\sqrt{\mu}}$  است. با افزایش  $\mu$ ، توزیع

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشاندیش

به سمت نرمال نزدیک می‌شود.

توزیع پواسون، هنگامی مورد استفاده قرار می‌گیرد که تعدادی پیشامد به طور تصادفی در طول زمان یا مکان رخ دهد، چنان‌که پیشامدها در زمان‌ها یا مکان‌های مجزا مستقل باشند. این توزیع همچنین به عنوان تقریبی برای توزیع دوجمله‌ای، زمانی که  $n$  بزرگ و  $\pi$  کوچک است ( $n\pi = \mu$ )، به کار می‌رود. یک مشخصه‌ی برجسته‌ی توزیع پواسون این است که واریانس اش برابر با میانگین اش است.

اجازه دهید به ارتباط بین توزیع‌های چندجمله‌ای و پواسون بپردازیم. فرض کنید  $Y_1, Y_2, Y_3$  تعداد افرادی باشد که در سانحه‌ی اتومبیل می‌میرند،  $\mu_1, \mu_2, \mu_3$  تعداد افرادی باشد که در سانحه‌ی هواپیمایی می‌میرند و  $\mu_4$  تعداد افرادی باشد که در سانحه‌ی قطار می‌میرند. یک مدل پواسون را در نظر بگیرید که در آن  $(Y_1, Y_2, Y_3)$  به‌گونه‌ای است که متغیرهای تصادفی  $Y_1$  و  $Y_2$  و  $Y_3$  مستقل و دارای توزیع پواسون، به ترتیب با پارامترهای  $\mu_1$  و  $\mu_2$  و  $\mu_3$  هستند. بنابراین تابع جرم احتمال برای  $Y_i$ ‌ها حاصل ضرب سه تابع جرم احتمال پواسون است. متغیر تصادفی  $\sum Y_i = n$  نیز دارای توزیع پواسون با پارامتر  $\sum \mu_i$  است.

با در نظر گرفتن نمونه‌گیری پواسون،  $N$  که مجموع شمارش‌ها است، به جای ثابت بودن، تصادفی است.  $Y_i$ ‌ها به شرط  $N$  مستقل نیستند، چرا که مقدار یکی روی مقداری که متغیرهای دیگر می‌گیرند، تأثیر دارد. برای  $c$  متغیر پواسون مستقل با  $\mu_i = \pi_i$ ، احتمال شرطی یک مجموعه از شمارش‌های  $n_i$  که در شرط  $N = \sum_{i=1}^c Y_i$  صدق کند به صورت زیر است:

$$\begin{aligned} \Pr[Y_1 = n_1, \dots, Y_c = n_c | \sum_{i=1}^c Y_i = N] &= \frac{\Pr(Y_1 = n_1, Y_2 = n_2, \dots, Y_c = n_c)}{\Pr(\sum Y_i = N)} \\ &= \frac{\prod_i \frac{\exp(-\mu_i) \mu_i^{n_i}}{n_i!}}{\frac{\exp(-\sum \mu_j) (\sum \mu_j)^N}{N!}} = \frac{N!}{\prod_i n_i!} \prod_i \pi_i^{n_i}, \end{aligned}$$

که در آن  $\sum_i \frac{\mu_i}{\mu_j} = \pi_i$  است. این یک توزیع چندجمله‌ای با پارامترهای  $(N, \pi)$  است که با اندازه‌ی بردار نمونه‌ی  $N$  و بردار احتمال  $(\pi_1, \dots, \pi_c) = \pi$  مشخص می‌شود. بسیاری از تحلیل‌های داده‌های رسته‌ای بر اساس توزیع چندجمله‌ای است. چنین تحلیل‌هایی معمولاً برآوردهای پارامتری مشابهی همانند تحلیل‌های بر اساس توزیع پواسون را به دلیل شباهت تابع‌های درست‌نمایی دارند.

## ۲.۵ مدل‌های لگ خطی برای جدول‌های پیشایندی

### دوطرفه

یک جدول پیشایندی  $J \times I$  از  $n$  آزمودنی را در نظر می‌گیریم که دو متغیر پاسخ  $X$  و  $Y$  را رده‌بندی می‌کند. فرض کنید شمارش‌های خانه‌ها،  $n_{ij}$ ، دارای توزیع پواسون با میانگین‌های  $\lambda_{ij}$  هستند. از آن جا که توزیع پواسون، عضوی از خانواده توزیع‌های نمایی با پارامتر طبیعی لگاریتم میانگین آن است، می‌توانیم از مدل لگ خطی زیر برای یافتن تأثیر سطوح مختلف دو متغیر  $X$  و  $Y$  بر تعداد شمارش‌های خانه‌ها استفاده کنیم.

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (1.5)$$

در این مدل،  $\lambda_i^X$  تأثیر سطر  $i$  از  $X$  است،  $\lambda_j^Y$  تأثیر ستون  $j$  از  $Y$  است، و  $\lambda_{ij}^{XY}$  تأثیر متقابل سطرهای  $i$  از  $X$  و  $j$  از  $Y$  بر لگاریتم تعداد شمارش‌ها را نشان می‌دهد.  $\lambda_{ij}^{XY}$  احراز از فرض استقلال دو متغیر را نشان می‌دهند. اگر  $\lambda_{ij}^{XY}$  به ازای هر  $i$  و هر  $j$  برابر با صفر باشد، مدل استقلال به دست می‌آید؛ به این معنا که

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y, \quad (2.5)$$

واز آن  $\mu_{ij} = (\exp \lambda)(\exp \lambda_i^X)(\exp \lambda_j^Y)$ ، که نشان می‌دهد یک مدل ضربی برای تعداد شمارش‌ها برقرار است. توجه کنید که در حالت استقلال داریم  $\mu_{ij} = p_{ij} = p_{i+}p_{+j}$ ، که از آن می‌توان نتیجه گرفت:

$$\mu_{ij} = np_{i+}p_{+j} = \frac{1}{n} \mu_{i+} \mu_{+j}$$

و با مقایسه با مدل ضربی (۲.۵) می‌توان به استقلال  $X$  و  $Y$  پی برد. برای شناساً بودن مدل، برخی محدودیت‌ها را باید بر پارامترهای مدل (۱.۵) اعمال کرد. این محدودیت‌ها را می‌توان به صورت‌های مختلف اعمال کرد. در اینجا فرض می‌کنیم که  $\lambda_I^X = 0$ ،  $\lambda_J^Y = 0$ ،  $\lambda_{ij}^{XY} = 0$  به ازای هر  $i$  و  $j$ . بنا بر این تعداد پارامترها در (۱.۵) عبارت است از  $IJ = (I - 1 + (I - 1)) + (J - 1 + (J - 1)) = 1 + (I - 1) + (J - 1) + (I - 1)(J - 1)$  که با تعداد خانه‌های جدول برابر است. به همین دلیل، این مدل را یک مدل اشباع‌شده می‌نامند. این مدل، کلی‌ترین مدل ممکن است که می‌توان برای یک جدول  $J \times I$  نوشت.

## ۱.۲.۵ تعبیر پارامترهای یک مدل دوطرفه

برای درک تعبیر پارامترها، یک جدول پیش‌بینی  $2 \times 2$  با مدل (۱.۵) را در نظر بگیرید. در این حالت، چهار پارامتر در مدل وجود دارد. با استفاده از مدل (۱.۵)، محدودیت‌های ذکر شده و لگاریتم نسبت بخت‌ها،  $(\log \theta)$ ، داریم:

$$\begin{aligned}\log \theta &= \log \left( \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} \right) = \log \mu_{11} + \log \mu_{22} - \log \mu_{12} - \log \mu_{21} \\ &= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) \\ &\quad - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY} = \lambda_{11}^{XY}\end{aligned}$$

که در آن تساوی آخر به علت اعمال محدودیت‌ها به دست آمده است. بنا بر این، فرض استقلال دو متغیر پذیرفته می‌شود اگر  $\lambda_{11}^{XY} = 0$  باشد.

حال اگر استقلال پذیرفته شود، در این جدول پیش‌بینی  $2 \times 2$  داریم:

$$\begin{aligned}\text{logit}[\Pr(Y = 1 | X = i)] &= \log \left[ \frac{\Pr(Y = 1 | X = i)}{\Pr(Y = 2 | X = i)} \right] = \log \left[ \frac{\Pr(Y = 1, X = i)}{\Pr(Y = 2, X = i)} \right] \\ &= \log \frac{\mu_{i1}}{\mu_{i2}} = \log \mu_{i1} - \log \mu_{i2} \\ &= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) = \lambda_1^Y - \lambda_2^Y = \lambda_1^Y.\end{aligned}$$

بنا بر این، احتمال موفقیت پاسخ  $Y$  به سطوح مختلف پاسخ  $X$  وابسته نیست و در نتیجه استقلال به این معنا است که  $\text{logit}[\Pr(Y = 1 | X = i)] = \alpha$ ، که در آن  $\alpha$  به وابسته نیست. می‌توان این را به عنوان مثال، این‌گونه تعبیر کرد که در تأثیر دو دارو بر روی بیماری سردرد، تفاوتی در درمان سردرد با استفاده از این دو دارو وجود ندارد. توجه کنید که مدل (۱.۵) یک مدل سلسله‌مراتبی نامیده می‌شود، به این معنا که اگر اثرهای مراتب بالاتر متغیرها در مدل موجود باشند (در اینجا اثرهای متقابل دوطرفه  $X$  و  $Y$ ) اثرهای پایین‌تر متغیرها (در اینجا اثرهای اصلی متغیرهای  $X$  و  $Y$ ) نیز در مدل وجود دارند. مدل‌های غیر سلسله‌مراتبی نیز در برخی کاربردها رخدان می‌دهند. مثالی از این مدل‌ها، مدل

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_{ij}^{XY}$$

است که در آن اثر متقابل  $X$  و  $Y$  در مدل موجود است، ولی اثر اصلی  $Y$  در مدل وجود ندارد. بهتر است برای تعبیر چنین مدل‌هایی به تعبیر اثرهای مراتب بالاترا اکتفا کنیم.

## ۲.۲.۵ انواع دیگر اعمال محدودیت‌ها

اعمال محدودیت‌ها در مدل (۱.۵) می‌تواند به صورت‌های دیگری نیز انجام شود. دیدیم که می‌توان سطح آخر هر متغیر در تأثیر اصلی آن را صفر فرض کرد. به هر حال می‌توان در نوعی دیگر از اعمال محدودیت‌ها، سطح اول متغیر را صفر فرض کرد یا می‌توان مجموع مؤلفه‌های پارامتری یک متغیر خاص را صفر فرض کرد. در حالت آخر به این محدودیت، محدودیت مجموع صفر اطلاق می‌شود. به هر صورت، هر گونه اعمالی برای محدودیت‌ها اختیار شود، مقادیر برازنده شده توسط مدل، یعنی  $z_{ijk}$ ‌ها برابر خواهند بود.

## ۳.۵ مدل‌های لگ خطی برای جدول‌های پیش‌ایندی سه‌طرفه

کلی‌ترین مدل لگ خطی برای یک جدول سه‌طرفه که در آن متغیرهای  $X$  با  $I$  سطح،  $Y$  با  $J$  سطح و  $Z$  با  $K$  سطح مدل‌بندی می‌شود، عبارت است از

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \quad (3.5)$$

که با اعمال محدودیت‌های  $\lambda_{ij}^{XY} = 0$ ،  $\lambda_k^Z = 0$ ،  $\lambda_j^Y = 0$ ،  $\lambda_I^X = 0$  به ازای هر  $j$ ،  $\lambda_{ij}^{XY} = 0$  به ازای هر  $i$ ،  $\lambda_{ik}^{XZ} = 0$  به ازای هر  $k$ ،  $\lambda_{jk}^{YZ} = 0$  به ازای هر  $j$  و هر  $k$ ،  $\lambda_{ijk}^{XYZ} = 0$  به ازای هر  $i$  و هر  $j$  و هر  $k$ ، تعداد پارامترهای این مدل عبارت است از

$$1 + (I - 1) + (J - 1) + (K - 1) + (I - 1)(J - 1) + (I - 1)(K - 1) \\ + (J - 1)(K - 1) + (I - 1)(J - 1)(K - 1) = IJK,$$

که چون تعداد پارامترها با تعداد خانه‌های جدول پیش‌ایندی سه‌طرفه یکسان است، یک مدل اشباع شده می‌باشد. در مدل (۳.۵)، همه‌ی اثرهای اصلی، اثرهای متقابل دوتایی و اثرهای متقابل سه‌تایی وجود دارند. نمادی که از آن برای معرفی این مدل

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشاندزی

استفاده می‌شود، نماد  $(XYZ)$  است. برای تمایز این نماد و این مدل با مدل‌هایی دیگر که  
حالت‌های خاص مدل  $(3.5)$  هستند، در ابتدا حالتی را در نظر بگیرید که  $\lambda_{ijk}^{XYZ} = 0$   
(بهارای همه‌ی  $i$ ها و  $j$ ها و  $k$ ها). در این صورت، مدل، همه‌ی اثرهای متقابل دوتایی و  
اثرهای اصلی را در برابر دارد، ولی همه‌ی اثرهای سه‌تایی، صفر هستند. در این صورت  
نیز مانند مدل اشباع‌شده‌ی  $(3.5)$ ، همه‌ی متغیرها به هم وابسته‌اند. این مدل را با نماد  
 $(XY, XZ, YZ)$  نمایش می‌دهند. اگر علاوه بر اثرهای متقابل سه‌تایی، اثر متقابل دوتایی  
دو تا از متغیرها نیز صفر باشد، برای تفسیر مدل، نیاز به اطلاعات در مورد مفهوم استقلال  
شرطی داریم. سه حالت است که علاوه بر اثرهای سه‌تایی، یکی از اثرهای متقابل دوتایی  
می‌تواند صفر باشد. حالت  $\lambda_{ijk}^{XYZ} = 0$  (بهارای همه‌ی  $i$ ها و  $j$ ها و  $k$ ها) و  $\lambda_{jk}^{YZ} = 0$   
(بهارای همه‌ی  $j$ ها و  $k$ ها) را در نظر می‌گیریم. بحث‌های زیر در مورد دو حالت دیگر  
می‌تواند به طور مشابه انجام شود. در این حالت، مدل عبارت است از

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}. \quad (4.5)$$

در این حالت، تعداد پارامترها به  $(1 - 1)(J - 1)(K - 1) - (I - 1)(J - 1)(K - 1) - (I - 1)(J - 1)$  تقلیل می‌یابد. همچنین  $Y$  و  $Z$  به شرط  $X$  از هم مستقل‌اند. به بیان دیگر،

$$\Pr(Y = j, Z = k | X = i) = \Pr(Y = j | X = i) \Pr(Z = k | X = i).$$

یا از دید احتمال شرطی و علائم مورد استفاده در این کتاب،

$$\pi_{jkl|i} = \pi_{j+i} \pi_{k+i}.$$

در این صورت برای احتمال‌های مشترک سه متغیر داریم:

$$\pi_{ijk} = \frac{\pi_{ij} + \pi_{ik}}{\pi_{i++}},$$

یا بر حسب میانگین شمارش‌های داخل جدول،

$$\mu_{ijk} = \frac{\mu_{ij} + \mu_{ik}}{\mu_{i++}}.$$

این مدل را با نماد  $(XY, XZ)$  نمایش می‌دهند.

اکنون حالتی را در نظر می‌گیریم که علاوه بر اثرهای متقابل سه‌تایی، دو مؤلفه از اثرهای

۳.۵ مدل‌های لگ خطی برای جداول‌های پیشانیدی سه‌طرفه

۱۳۷

متقابل دوتایی نیز صفر باشند. حالت زیر را در نظر بگیرید:

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}, \quad (5.5)$$

که در آن به ازای هر  $i$  و  $j$  و  $k$  داریم  $\lambda_{ijk}^{XYZ} = 0$  و  $\lambda_{ik}^{XZ} = 0$  و  $\lambda_{jk}^{YZ} = 0$  و  $\lambda_{ij}^{XY} = 0$ . (بحث در مورد دو حالت دیگر، مشابه بحث در مورد مدل ۵.۵ است). در این حالت، تعداد پارامترها عبارت است از

$$1 + (I - 1) + (J - 1) + (K - 1) + (I - 1)(J - 1).$$

در این حالت،  $X$  و  $Y$  مشترکاً از  $Z$  مستقل‌اند؛ یعنی

$$\Pr(X = i, Y = j, Z = k) = \Pr(X = i, Y = j) \Pr(Z = k).$$

به بیان دیگر،

$$\pi_{ijk} = \pi_{ij+} \pi_{++k},$$

واز آن‌جا

$$\mu_{ijk} = \frac{\mu_{ij+} \mu_{++k}}{n}.$$

این مدل را با نماد  $(XY, Z)$  نمایش می‌دهند.

مدل زیر را مدل استقلال می‌نامند:

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z,$$

در این مدل که با نماد  $(X, Y, Z)$  نمایش داده می‌شود،  $X$  و  $Y$  و  $Z$  دو به دو مستقل‌اند و داریم:

$$\pi_{ijk} = \pi_{i++} \pi_{+j+} \pi_{++k},$$

که در آن،

$$\mu_{ijk} = \frac{\mu_{i++} \mu_{+j+} \mu_{++k}}{n^3}.$$

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیش‌اندی

### ۱.۳.۵ تعبیر پارامترهای یک مدل سه‌طرفه

برای تعبیر پارامترها در مدل لگ خطی برای سه متغیر، باید بالاترین مرتبه‌ی اثرهای متقابل را در نظر گرفت. برای این کار، نیاز به تعبیر پیوند شرطی است. پیوند شرطی بین  $X$  و  $Y$ ، از همه‌ی نسبت‌های بخت استفاده می‌کند. نسبت‌های بخت موضعی شرطی بین  $X$  و  $Y$ ، به شرط  $k$  امین سطح  $Z$ ، برابرند با

$$\theta_{ij(k)} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\pi_{i+1,j,k}}, \quad 1 \leq i \leq I-1, \quad 1 \leq j \leq J-1$$

که  $(I-1)(J-1)$  تا از چنین نسبت‌های بختی وجود دارند. به طور مشابه  $\theta_{i(j)k}$ ، تعداد  $(I-1)(K-1)$  نسبت بخت شرطی بین  $Y$  و  $Z$  را به شرط  $X$  تعریف می‌کند. استقلال شرطی  $X$  و  $Y$ ، هم‌ارز آن است که به‌ازای هر  $i$  و  $j$  و  $k$  داشته باشیم  $\theta_{ij(k)} = 1$ . پارامترهای اثرهای متقابل با نسبت‌های بخت شرطی در ارتباط‌اند. به عنوان مثال، برای مدل  $(XY, XZ, YZ)$  داریم:

$$\begin{aligned} \log \theta_{ij(k)} &= \log \mu_{ijk} + \log \mu_{i+1,j+1,k} - \log \mu_{i+1,j,k} - \log \mu_{i,j+1,k} \\ &= \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}. \end{aligned} \quad (6.5)$$

از آن‌جا که  $\log \theta_{ij(k)}$  مرتبط با  $k$  در سمت راست معادله‌ی (6.5) نیست، برای مدل  $(XY, XZ, YZ)$  داریم:

$$\theta_{ij(1)} = \theta_{ij(2)} = \cdots = \theta_{ij(K)}, \quad \forall i, j$$

همچنین می‌توان نشان داد که

$$\theta_{i(1)k} = \theta_{i(2)k} = \cdots = \theta_{i(J)k}, \quad \forall i, k$$

و

$$\theta_{(1)jk} = \theta_{(2)jk} = \cdots = \theta_{(I)jk}, \quad \forall j, k$$

این است دلیل آن که چرا می‌گوییم در مدل  $(XY, XZ, YZ)$  پیوند همگن وجود دارد.

## ۴.۵ مدل‌های لگ خطی برای جدول‌های دارای مراتب بالاتر

برای جدول‌های مراتب بالاتر از سه‌طرفه (به عنوان مثال، چهار‌طرفه) به طریقی مشابه می‌توان مدل‌های لگ خطی را تعریف کرد. به عنوان مثال، مدل  $(X, Y, Z, W)$  را که به صورت زیر است، در نظر بگیرید.

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{kl}^{ZW},$$

که در آن محدودیت‌ها عبارت‌اند از  $\lambda = 0$ ,  $\lambda_I^X = 0$ ,  $\lambda_J^Y = 0$ ,  $\lambda_K^Z = 0$ ,  $\lambda_{kl}^{ZW} = 0$  به ازای هر  $i, j, k, l$ . در این مدل،  $Z$  و  $W$  مشترکاً از  $X$  و  $Y$  مستقل‌اند.تابع درست‌نمایی و آماره‌های بسندۀ می‌توانند به روشی مشابه با جدول‌های سه‌طرفه یافته و تعریف شوند (بخش ۶.۵ را ببینید).

## ۵.۵ نیکویی برازش مدل لگ خطی

برای مقایسه‌ی عملکرد دو مدل مختلف می‌توان از آماره‌های  $\chi^2$  و  $G^2$  (به ترتیب، آماره‌ی آزمون خی دوپی‌یرسون و آزمون نسبت درست‌نمایی تعمیم‌یافته) استفاده کرد. معیار دیگری که می‌توان از آن استفاده کرد، معیار اطلاع آکائیکه است که به صورت

$$AIC = -2[G^2 + 2(d.f.)] \quad (\text{تعداد پارامترهای مدل}) - (\text{درست‌نمایی ماکسیمم})$$

تعریف می‌شود. هرچه  $AIC$  کوچک‌تر باشد، مدل دارای برازش بهتری است. به عنوان مثال، برای مقایسه‌ی مدل اشباع‌شده ( $M_1$ ) با مدلی که در آن همه‌ی اثرهای متقابل سه‌تایی برابر صفرند ( $M_0$ ), می‌توان از  $(L_{M_0} - L_{M_1})/2$  استفاده کرد که در آن  $L_{M_0}$  لگاریتم درست‌نمایی تحت مدل  $M_0$  است و  $L_{M_1}$  لگاریتم درست‌نمایی تحت مدل  $M_1$ . این معادل است با

$$G^2(M_0|M_1) = G^2(M_0) - G^2(M_1),$$

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیش‌اندی

که در آن  $G^2(M_0)$  کیبیش مدل  $M_0$  است و  $G^2(M_1)$  کیبیش مدل  $M_1$  است.  
 $G^2(M_0|M_1)$  تحت مدل  $M_0$  دارای توزیع مجانبی خی دو با درجهٔ آزادی

$$\text{d.f.} = (M_1) - (\text{تعداد پارامترهای مدل } M_0)$$

است که در این مثال، برابر با  $(I-1)(J-1)(K-1)$  است. مدل‌های دیگر به روش مشابه می‌توانند مورد مقایسه قرار گیرند.

## ۶.۵ روش‌های براورد

برای براورد کردن پارامترهای یک مدل لگ خطی، از روش‌های گوناگونی می‌توان استفاده کرد. در ابتدا روش ماکسیمم درست‌نمایی با به دست آوردن معادلات درست‌نمایی تشریح می‌شود. سپس روش برآش مناسب تکراری مورد بحث قرار می‌گیرد.

## ۱.۶.۵ روش ماکسیمم درست‌نمایی

در ابتدا یک مدل اشباع‌شده برای جدول‌های سه‌طرفه را در نظر بگیرید. فرض می‌کنیم که شمارش‌های خانه‌ها دارای توزیع پواسون باشند. در این صورت، تابع درست‌نمایی عبارت است از

$$L = \prod_{i=1}^I \prod_{j=1}^J \prod_{k=1}^K \frac{e^{\mu_{ijk}} (\mu_{ijk})^{n_{ijk}}}{n_{ijk}!}.$$

هستهٔ لگاریتم این تابع درست‌نمایی عبارت است از

$$L(\mu) = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ijk} \log \mu_{ijk} - \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \mu_{ijk}.$$

برای مدل اشباع‌شدهٔ (۳.۵) داریم:

$$L(\mu) = n\lambda + \sum_{i=1}^I n_{i++} \lambda_i^X + \sum_{j=1}^J n_{+j+} \lambda_j^Y + \sum_{k=1}^K n_{++k} \lambda_k^Z$$

$$\begin{aligned}
 & + \sum_{i=1}^I \sum_{j=1}^J n_{ij} + \lambda_{ij}^{XY} + \sum_{i=1}^I \sum_{k=1}^K n_{i+k} \lambda_{ik}^{XZ} + \sum_{j=1}^J \sum_{k=1}^K n_{+jk} \lambda_{jk}^{YZ} \\
 & + \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ijk} \lambda_{ijk}^{XYZ} - \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \exp(\lambda + \dots + \lambda_{ijk}^{XYZ}). \quad (7.5)
 \end{aligned}$$

چون توزیع پواسون از خانواده‌ی نمایی است، ضرایب پارامترها آماره‌های بستنده‌ی کمینه هستند. در مدل اشباع شده،  $n_{ijk} \lambda_{ijk}^{XYZ}$  ضرایب هستند و بنا بر این کاهشی در مشاهدات توسط آماره‌ی بستنده دیده نمی‌شود. برای مدل‌های ساده‌تر، به عنوان مثال برای مدل  $(XY, XZ, YZ)$ ، در معادله‌ی درست‌نمایی،  $\lambda_{ijk}^{XYZ}$ ‌ها برابر با صفرند و در این صورت، آماره‌های بستنده عبارت‌اند از  $n_{i+k}$  و  $n_{+jk}$  و  $n_{ij}$ . برای مدل‌های ساده‌تر دیگر، آماره‌های بستنده می‌توانند به روشی مشابه به دست آیند. برای به دست آوردن معادلات درست‌نمایی، اجازه دهید تا درست‌نمایی را به صورت ماتریسی بنویسیم.

اگر  $'n = (n_1, \dots, n_N)$  و  $'\mu = (\mu_1, \dots, \mu_N)$  به ترتیب، بردارهای ستونی مشاهدات و مقادیر مورد انتظار  $N$  خانه‌ی یک جدول پیش‌اندی را نشان دهند، مدل‌های لگ‌خطی برای میانگین‌های پواسون دارای صورت زیرند:

$$\log \mu = X'\beta, \quad (8.5)$$

که در آن  $X$  ماتریس طرح و  $\beta$  بردار ستونی پارامترهای مدل است. به عنوان مثال، برای مدل  $I = J = K = 2$  با  $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$  و محدودیت‌های آخرین سطر صفر، داریم:

$$\log \begin{bmatrix} \mu_{111} \\ \mu_{112} \\ \mu_{121} \\ \mu_{211} \\ \mu_{221} \\ \mu_{212} \\ \mu_{122} \\ \mu_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} \lambda \\ \lambda_1^X \\ \lambda_1^Y \\ \lambda_1^Z \end{bmatrix}.$$

فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشاندزی

برای مدل عمومی (۷.۵)، لگاریتم درست‌نمایی عبارت است از

$$\begin{aligned}\ell(\mu) &= \log L(\mu) = \sum_i n_i \log \mu_i - \sum_i \mu_i \\ &= \sum_i n_i \left( \sum_j x_{ij} \beta_j \right) - \sum_i \exp \left( \sum_j x_{ij} \beta_j \right).\end{aligned}$$

آماره‌ی بسنده برای  $\beta_j$ ، برابر است با  $\sum_{i=1} n_i x_{ij}$  و داریم:

$$\frac{\partial}{\partial \beta_j} \left[ \exp \left( \sum_j x_{ij} \beta_j \right) \right] = x_{ij} \exp \left( \sum_j x_{ij} \beta_j \right) = x_{ij} \mu_i,$$

$$\frac{\partial \ell(\mu)}{\partial \beta_j} = \sum_i n_i x_{ij} - \sum_i \mu_i x_{ij}, \quad j = 1, 2, \dots, P$$

که در آن  $P$  تعداد پارامترهای مدل است.

بنا براین، معادلات درست‌نمایی عبارت‌اند از

$$X' n = X' \hat{\mu}. \quad (9.5)$$

همان‌طور که این معادلات نشان می‌دهند، آماره‌های بسنده در این معادلات با مقادیر مورد انتظارشان مساوی قرار داده می‌شوند. به عنوان مثال، برای مدل  $(XZ, YZ)$  داریم:

$$\begin{aligned}\frac{\partial L}{\partial \lambda_{ik}^{XZ}} &= n_{i+k} - \mu_{i+k}, \\ \frac{\partial L}{\partial \lambda_{jk}^{YZ}} &= n_{+jk} - \mu_{+jk},\end{aligned}$$

که به ترتیب به معادلات زیر منجر می‌شوند:

$$\hat{\mu}_{i+k} = n_{i+k}, \quad \forall i, k$$

$$\hat{\mu}_{+jk} = n_{+jk}, \quad \forall j, k$$

برای سایر مدل‌ها می‌توان معادلات درست‌نمایی را به‌طور مشابه حدس زد. به عنوان مثال، برای  $(X, Y, Z)$  داریم:

$$\hat{\mu}_{i++} = n_{i++}, \quad \forall i$$

$$\hat{\mu}_{+j+} = n_{+j+}, \quad \forall j$$

$$\hat{\mu}_{++k} = n_{++k}, \quad \forall k$$

ماتریس‌های کوواریانس براوردهای ML می‌توانند با مشتق‌گیری دوباره‌ی تابع لگاریتم درست‌نمایی به دست آیند، که بر اساس آن داریم:

$$\begin{aligned}\frac{\partial^2 L(\mu)}{\partial \beta_j \partial \beta_k} &= \sum_i x_{ij} \frac{\partial \mu_i}{\partial \beta_k} \\ &= - \sum_i x_{ij} \left( \frac{\partial}{\partial \beta_k} \exp\left(\sum_h x_{ih} \beta_h\right) \right) = - \sum_i x_{ij} x_{ik} \mu_i,\end{aligned}$$

و بنا بر این، ماتریس اطلاع عبارت است از

$$I = X' \text{diag}(\mu) X$$

که در آن  $\text{diag}(\mu)$  دارای دایه‌های  $\mu$  روی قطر اصلی است. برای نمونه‌های بزرگ،  $\hat{\beta}$  دارای توزیع مجانبی نرمال با میانگین  $\beta$  و ماتریس کوواریانس  $I^{-1}$  است. بنا بر این،

$$\text{cov}(\hat{\beta}) = [X' \text{diag}(\mu) X]^{-1}.$$

برای مدل‌های لگ خطی که برای سه متغیر  $X$  و  $Y$  و  $Z$  مورد بحث قرار دادیم، یافتن براورد مقادیر برازانده شده‌ی  $\hat{\mu}_{ijk}$  برای مقایسه‌ی مدل‌های مختلف (با استفاده از  $\chi^2$  یا  $G^2$ ) کافی است (نیازی به براورد پارامترها نداریم). به هر حال در همه‌ی مدل‌های بحث شده نمی‌توان مقادیر برازانده شده را به صورت بسته یافت. به عنوان مثال، برای مدل  $(XYZ)$  مقادیر برازانده شده عبارت‌اند از  $\hat{\mu}_{ijk} = n_{ijk}$  که از آن  $\chi^2 = G^2 = \sum_{ijk} (\hat{\mu}_{ijk} - \mu_{ijk})^2 / \mu_{ijk}$  به دست می‌آید.

مثال: برای مدل  $(XY, XZ, YZ)$  صورت بسته‌ای برای  $\hat{\mu}_{ijk}$  وجود ندارد و فقط با روش‌های تکراری می‌توان  $\hat{\mu}_{ijk}$  را به دست آورد. برای سایر مدل‌های جدول‌های سه‌طرفه، صورت بسته برای  $\hat{\mu}_{ijk}$  وجود دارد. به عنوان مثال، برای مدل  $(XZ, YZ)$  از آن جا که  $X$  و  $Y$  به شرط  $Z$  مستقل‌اند، داریم:

$$\pi_{ijk} = \frac{\pi_{i+k} \pi_{+jk}}{\pi_{++k}}, \quad \forall i, j, k$$

و بنا بر این:

$$\hat{\mu}_{ijk} = \frac{\hat{\mu}_{i+k} \hat{\mu}_{+jk}}{\hat{\mu}_{++k}} = \frac{n_{i+k} n_{+jk}}{n_{++k}}.$$

مثال: برای مدل  $(X)$  داریم:

$$\mu_{ijk} = \lambda + \lambda_i^X,$$

واز آن جا  $\pi_{ijk} = C\pi_{i++}$  که در آن  $C$  یک مقدار ثابت است. اما داریم  $C = \frac{1}{JK}$  که نتیجه می‌دهد  $1 = \sum_i \sum_j \sum_k \pi_{ijk}$  و در نتیجه  $\hat{\mu}_{ijk} = \frac{n_{i++}}{JK}$  واز آن جا  $\pi_{ijk} = \frac{\pi_{i++}}{JK}$

### ۱.۱.۶.۵ نتیجه‌ی برج برای مدل‌های لگ خطی

برج (۱۹۶۳) نشان داد که معادلات درست‌نمایی برای مدل‌های لگ خطی، آماره‌های بسنده‌ی کمینه را با مقادیر امید ریاضی این آماره‌ها تطبیق می‌دهند؛ به این معنا که اگر مقادیر برازانده‌شده‌ای معرفی شوند که در مدل صدق کنند و داده‌ها را با آماره‌های بسنده‌ی کمینه تطابق دهند، این مقادیر برازانده‌شده، براوردهای ML خواهند بود.

مثال: مدل  $(X, Y, Z)$  را برای یک جدول سه‌طرفه در نظر بگیرید. داریم:

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z,$$

که آماره‌های بسنده‌ی کمینه‌ی آن عبارت‌اند از  $\{n_{i++}\}$  و  $\{n_{j++}\}$  و  $\{n_{k++}\}$ . معادلات درست‌نمایی عبارت‌اند از

$$\hat{\mu}_{i++} = n_{i++}, \quad \forall i$$

$$\hat{\mu}_{j++} = n_{j++}, \quad \forall j$$

$$\hat{\mu}_{k++} = n_{k++}, \quad \forall k$$

مقادیر برازانده‌شده‌ی  $\hat{\mu}_{ijk} = \frac{n_{i++}n_{j++}n_{k++}}{n!}$  در این معادلات و نیز در مدل صدق می‌کنند. بنا براین طبق نتیجه‌ی برج، این براوردها براوردهای ML هستند.

## ۲.۶.۵ روش‌های تکراری برآش مدل لگخطی

### ۱.۲.۶.۵ روش نیوتن-رافسون

همان‌طور که قبل اشاره شد،تابع لگاریتم درست‌نمایی عبارت بود از

$$\sum_i n_i \left( \sum_h x_{ih} \beta_h \right) - \sum_i \exp\left(\sum_h x_{ih} \beta_h\right).$$

بنا براین

$$u_j = \frac{\partial L(\beta)}{\partial \beta_j} = \sum_i n_i x_{ij} - \sum_i \mu_i x_{ij},$$

$$h_{jk} = \frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} = - \sum_i \mu_i x_{ij} x_{ik}.$$

در نتیجه:

$$u_j^{(t)} = \sum_i (n_i - \mu_i^{(t)}) x_{ij},$$

و

$$h_{jk}^{(t)} = - \sum_i \mu_i^{(t)} x_{ij} x_{ik},$$

واز آن‌جا

$$\beta^{(t+1)} = \beta^{(t)} + [X' \text{diag}(\mu^{(t)}) X]^{-1} X' (n - \mu^{(t)}),$$

که  $\mu^{(t+1)}$  با استفاده از  $\mu^{(t+1)} = \exp(X\beta^{(t+1)})$  به دست می‌آید.  $\beta^{(t+1)}$  می‌تواند از رابطه‌ی زیر به دست آید:

$$\beta^{(t+1)} = -(H^{(t)})^{-1} r^{(t)},$$

که در آن  $r_j^{(t)} = \sum \mu_i^{(t)} x_{ij} [\log \mu_i^{(t)} + \frac{(n_i - \mu_i^{(t)})}{\mu_i^{(t)}}]$  نیز محاسبه می‌شود. به عنوان مقدار اولیه می‌توان از  $\mu_i^{(0)} = n_i + \frac{1}{\gamma}$  یا  $\mu_i^{(0)} = n_i$  استفاده کرد.

## ۲.۶.۵ برآش متناسب تکراری

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشاندیشی

الگوریتم برازش مناسب تکراری<sup>۱</sup> (IPF) روشی ساده برای محاسبهٔ  $\mu_{ijk}$  برای مدل‌های لگ خطی است. این روش توسط دمینگ و استفن (۱۹۴۰) معرفی شده است. گام‌های

این الگوریتم عبارت‌اند از

۱. با  $\{\mu_{ijk}^{(0)}\}$  (به عنوان مثال به‌ازای هر  $i$ ،  $j$ ،  $k$ )  $1/00 \equiv \mu_{ijk}^{(0)}$  شروع کنید.

۲. با ضرب عامل‌های مناسب،  $\{\mu_{ijk}^{(0)}\}$  را طوری تعديل کنید که مجموعهٔ آماره‌های بسنده‌ی کمینه با جدول‌های حاشیه‌ای جورسازی شوند.

۳. آنقدر ادامه دهید تا ماقسیم اختلاف بین آماره‌های بسنده و مقادیر برازنده شده به اندازه‌ی کافی به صفر نزدیک شوند.

این روش را با استفاده از تنها مدلی که در بحث‌های قبلی سه متغیره، صورت بسته‌ای برای  $n_{ijk}$  نداشت، یعنی مدل  $(XY, XZ, YZ)$ ، تشریح می‌کنیم.

در این مدل، آماره‌های بسنده‌ی کمینه عبارت‌اند از  $\{n_{ij+}\}$  و  $\{n_{i+k}\}$  و  $\{n_{++k}\}$ . برآوردهای اولیه باید در این مدل صدق کنند. اولین گام IPF به صورت زیر است:

$$\begin{aligned} \mu_{ijk}^{(1)} &= \mu_{ijk}^{(0)} \frac{n_{ij+}}{\mu_{ij+}^{(0)}}, \\ \mu_{ijk}^{(2)} &= \mu_{ijk}^{(1)} \frac{n_{i+k}}{\mu_{i+k}^{(1)}}, \\ \mu_{ijk}^{(3)} &= \mu_{ijk}^{(2)} \frac{n_{++k}}{\mu_{++k}^{(2)}}. \end{aligned} \quad (10.5)$$

اگر هر دو قسمت عبارت اول فرمول (۱۰.۵) را روی  $k$  جمع بیندیم، برای هر  $i$  و  $j$  داریم  $\mu_{ij+}^{(1)} = n_{ij+}$ . بنا بر این بعد از گام ۱، مقادیر مشاهده شده و برازنده شده در جدول حاشیه‌ای  $XY$  جورسازی می‌شوند. بعد از مرحله‌ی ۲، برای هر  $i$  و  $k$  داریم  $\mu_{i+k}^{(2)} = n_{i+k}$ ، اما جدول‌های حاشیه‌ای  $XY$  در شرط جورسازی صدق نمی‌کنند. بعد از مرحله‌ی ۳،  $\mu_{++k}^{(3)} = n_{++k}$  است؛ اما جدول‌های حاشیه‌ای  $XY$  و  $XZ$  دیگر در شرط جورسازی صدق نمی‌کنند. بنا بر این یک چرخه‌ی دیگر الگوریتم را تکرار می‌کنیم و  $\mu_{ijk}^{(3)}$  را محاسبه می‌کنیم و به همین ترتیب ادامه می‌دهیم. این کار تا آن‌جا ادامه می‌یابد که ماقسیم اختلاف بین آماره‌های بسنده و مقادیر برازنده شده کمتر از مقدار معلوم از پیش داده شده پیش رود.

## ۷.۵ مثال کاربردی

### ۱.۷.۵ داده‌های مربوط به تصادفات در ایالت میان آمریکا

با استفاده از مثال زیر که از اگرستی (۲۰۰۲) گرفته شده است، روش تحلیل مدل‌های لگ خطی را تشریح می‌کنیم. جدول (۱.۵) مشاهدات ۶۸۶۹۴ مسافر را نشان می‌دهد که در ایالت میان آمریکا در سال ۱۹۹۱ میلادی تصادف داشته‌اند. این جدول بر حسب جنس (G)، محل تصادف (L)، استفاده از کمربند ایمنی (S) و آسیب‌دیدگی (I) رده‌بندی شده است.

جدول ۱.۵: داده‌های مربوط به تصادفات در ایالت میان آمریکا (منبع: اگرستی، ۲۰۰۲)

نسبت نمونه	(GLS, GI, IL, IS)		(GI, GL, GS, IL, LS)		آسیب دیدگی		جنس	مکان	زن
	داشته	نداشته	داشته	نداشته	داشته	نداشته			
۰/۱۲	۱۰۰۹/۸	۷۲۲۲/۲	۹۹۳/۰	۷۱۶۶/۴	۹۹۶	۷۲۸۷	بسته	شهری	
۰/۰۶	۲۱۲/۴	۱۱۶۳۲/۶	۷۲۱/۳	۱۱۷۴۸/۳	۷۵۹	۱۱۵۸۷	بسته		
۰/۲۳	۱۱۴/۳	۲۲۵۴/۲	۹۸۸/۸	۳۲۵۳/۸	۹۷۲	۳۲۴۶	بسته	روستایی	
۰/۱۱	۲۹۷/۵	۶۰۹۳/۵	۷۸۱/۹	۵۹۸۵/۵	۷۵۷	۶۱۳۴	بسته		
۰/۰۷	۸۲۴/۱	۱۰۳۵۸/۹	۸۴۵/۱	۱۰۴۲۱/۵	۸۱۲	۱۰۳۸۱	بسته	شهری	مرد
۰/۰۳	۲۸۹/۸	۱۰۱۰۹/۲	۳۸۷/۶	۱۰۸۳۷/۸	۳۸۰	۱۰۹۶۹	بسته		
۰/۱۵	۱۰۵۶/۸	۶۱۵۰/۲	۱۰۳۸/۱	۶۰۴۵/۳	۱۰۸۴	۶۱۲۳	بسته	روستایی	
۰/۰۲	۵۰۸/۴	۶۶۱۲/۶	۵۱۸/۲	۶۸۱۱/۴	۵۱۲	۶۶۹۳	بسته		

جدول (۲.۵) نتایج آزمون نیکویی برازش برای چند مدل لگ خطی را نشان می‌دهد. همان‌طور که نتایج جدول نشان می‌دهد، مدل استقلال دویه‌دو (G,I,L,S) خیلی ضعیف عمل می‌کند. مدل (GI,GL,GS,IL,IS,LS) خیلی بهتر برازانده شده است، ولی هنوز فقدان برازش مدل وجود دارد ( $P < 0.001$ —مقدار). مدل (GIL, GIS, GLS, ILS) دارای برازش خیلی خوبی است ( $G^2 = 1/3$ , d.f. = ۱,  $\chi^2 = 1/3$ ), اما خیلی پیچیده است و برای تعبیر، مدل مشکلی است. بنا براین به مدلی پیچیده‌تر از (GI,GL,GS,IL,IS,LS) (GIL, GIS, GLS, ILS) نیاز داریم. به هر حال، اگرستی (۲۰۰۲، ص. ۳۲۸) ساده‌تر از (GIL, GIS, GLS, ILS) ترجیح داده است تا از مدل (GI,GL,GS,IL,IS,LS) استفاده کند که روی اثرهای متقابل دوتایی تمرکز کرده است. جدول (۲.۵) مقادیر برازش‌یافته توسط این مدل را نشان می‌دهد. مقایسه‌ی دو مدل (GI,GL,GS,IL,IS,LS) و (GI, GL, GS, IL, IS, LS) اختلاف  $G^2$ -یی برابر با  $15/9 = 15 - 7/5 = ۲۳/۴ - ۴ = ۱$  درجه‌ی آزادی ( $P = 0.0001$ —مقدار) را نشان می‌دهد که حاکی از برتری عملکرد مدل (GLS, GI, IL, IS) از دیدگاه نظری است.

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشانندی

جدول ۲.۵: نتایج آزمون نیکویی برازش برای چند مدل لگ خطی

$-P$	d.f.	$G^2$	مدل
< ۰/۰۰۰۱	۱۱	۲۷۹۲/۸	(G,I,L,S)
< ۰/۰۰۱	۵	۲۳/۴	(GI,GL,GS,IL,IS,LS)
۰/۲۵	۱	۱/۳	(GIL,GIS,GLS,ILS)
۰/۰۰۱	۴	۱۸/۶	(GIL,GS,IS,LS)
< ۰/۰۰۱	۴	۲۲/۸	(GIS,GL,IL,LS)
۰/۱۱	۴	۷/۵	(GLS,GI,IL,IS)
< ۰/۰۰۱	۴	۲۰/۶	(ILS,GI,GL,GS)

اگرستی (۲۰۰۲، ص. ۳۲۹) از شاخص براورد عدم تشابه که به صورت زیر تعریف می‌شود، استفاده کرده است تا نشان دهد که اگرچه از دیدگاه نظری بین دو مدل ذکر شده اختلاف وجود دارد، از دیدگاه کاربردی، اختلاف فاحشی در عملکرد این دو مدل وجود ندارد.

$$\hat{\Delta} = \frac{\sum_i |n_i - \hat{\mu}_i|}{2n} = \frac{\sum_i |p_i - \hat{p}_i|}{2},$$

که در آن  $\hat{\Delta} = \Delta$  وقتی اتفاق می‌افتد که مدل خیلی خوب (در حد اعلا) عمل کرده باشد. وقتی  $\hat{\Delta} > \Delta$  یا  $\hat{\Delta} < \Delta$  باشد، نمونه از مدل، خیلی خوب پیروی کرده است، هرچند که مدل کامل نیست. برای مدل (GLS,GI,IL,IS),  $\hat{\Delta} = ۰/۰۰۸$  است و برای مدل (GI,GL,GS,IL,IS,LS),  $\hat{\Delta} = ۰/۰۰۳$  است که نشان می‌دهد هر دو مدل می‌توانند برازش‌های خوبی تلقی شوند.

## ۸.۵ دستورهای R برای برازش مدل‌های لگ خطی

در این بخش به تحلیل یک جدول  $2 \times 2$  می‌پردازیم که رنگ مو و چشم تعدادی مرد را رده‌بندی کرده است (جدول ۳.۵ را ببینید).

جدول ۳.۵: داده‌های مربوط به رنگ مو و چشم تعدادی مرد

رنگ چشم	رنگ مو
سبز	مشکی
۳	قهوه‌ای
۱۰	آبی
۱۵	قهوه‌ای
۲۵	آبی
۷	قرمز
۸	بور
۵	
۳۰	
۳	

## ۱.۵ دستورهای R برای برازش مدل‌های لگ خطی

۱۴۹

از دستورهای زیر برای ورود داده‌ها و چاپ داده‌ها استفاده می‌کنیم:

```
hair = factor(rep(c("Black", "Brown", "Red", "Blond"), c(4,4,4,4)))
eye  = factor(rep(c("Brown", "Blue", "Hazel", "Green"), 4))
freq = c(32,11,10,3, 38,50,25,15, 10,10,7,7, 3,30,5,8)
haireye.data <- data.frame(hair, eye, freq)
```

با استفاده از دستورهای زیر، داده‌ها داخل یک جدول قرار می‌گیرند و به صورت data frame در می‌آیند.

```
haireye.table = table(hair, eye)
haireye.table[cbind(hair,eye)] = freq
```

و با دستور زیر می‌توان جدول را مشاهده کرد.

```
haireye.table
```

توجه کنید که در این جدول، سطوح عامل به ترتیب حروف الفبا دوباره مرتب شده‌اند. اکنون اجازه دهید برای شروع، مدلی با محدودیت‌های مجموع صفر را برازش دهیم. دستور زیر، این محدودیت را اعمال می‌کند.

```
options(contrasts=c("contr.sum", "contr.poly"))
```

دستور زیر، مدل استقلال را با استفاده از دستور `glm` برازش می‌دهد.

```
glm.out1 = glm(freq~hair+eye, family=poisson, data=haireye.data)
summary(glm.out1)
```

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشانیدی

۱۵۰

```

Call:
glm(formula = freq ~ hair + eye, family = poisson, data = haireye.data)

Deviance Residuals:
    Min      1Q   Median      3Q     Max 
-3.6724 -0.9527 -0.1018  0.5635  3.0744 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 2.57730   0.07489 34.413 < 2e-16 ***
hair1       -0.03274   0.11717 -0.279  0.77990  
hair2       -0.22945   0.12518 -1.833  0.06681 .  
hair3        0.79393   0.09331  8.508 < 2e-16 ***
eye1        0.51997   0.09770  5.322 1.03e-07 ***
eye2        0.32369   0.10305  3.141  0.00168 ** 
eye3        -0.59865   0.14052 -4.260 2.04e-05 *** 
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 163.492 on 15 degrees of freedom
Residual deviance: 44.315 on 9 degrees of freedom
AIC: 127.46

Number of Fisher Scoring iterations: 5

```

شکل ۲.۵: نتایج استفاده از مدل استقلال با استفاده از دستور `glm`

خروجی این اجرا در شکل (۲.۵) داده شده است.

مقدار کیبس ۳۱۵/۴۴ با ۹ درجه‌ی آزادی نشان می‌دهد که باید این مدل رد شود.

برای محاسبه‌ی  $P$ -مقدار از دستور زیر استفاده می‌کنیم:

```
1-pchisq(44.315,9)
```

```
[1] 1.234755e-06
```

بنا بر این نیاز داریم تا مدل اشباع‌شده را برازش دهیم. دستور زیر، این کار را انجام می‌دهد، که نتایج آن در شکل (۲.۵) داده شده است.

```
glm.out2 <- glm(freq~hair*eye, family=poisson, data=haireye.data)
summary(glm.out2)
```

می‌توانیم مدل‌های بالا را با دستور `loglik` نیز برازش دهیم. در این صورت برای مدل استقلال از دستورهای زیر استفاده می‌کنیم و نتایج آن در شکل (۴.۵) آمده است.

```

Call:
glm(formula = freq ~ hair * eye, family = poisson, data = haireye,data)

Deviance Residuals:
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.464190  0.085829 28.710 < 2e-16 ***
hair1       -0.147983  0.157244 -0.941 0.346652
hair2       -0.417018  0.170239 -2.450 0.014301 *
hair3        0.904944  0.110215  0.211 < 2e-16 ***
eye1         0.539235  0.122471  4.403 1.07e-05 ***
eye2         0.161940  0.150689  1.075 0.282527
eye3        -0.506187  0.168039 -3.012 0.002593 **
hair1:eye1  -0.457547  0.234703 -1.949 0.051239 .
hair2:eye1  0.814790  0.212007  3.843 0.000121 ***
hair3:eye1  0.003654  0.157416  0.023 0.981480
hair1:eye2  0.987589  0.218812  4.513 6.38e-06 ***
hair2:eye2 -1.110500  0.357289 -3.108 0.001883 **
hair3:eye2  0.106513  0.184572  0.577 0.563886
hair1:eye3  -0.711408  0.359068 -1.981 0.047562 *
hair2:eye3  0.538456  0.284784  1.891 0.058657 .
hair3:eye3 -0.154897  0.222900 -0.695 0.487108

---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

Dispersion parameter for poisson family taken to be 1

Null deviance: 1.6349e+02 on 15 degrees of freedom
Residual deviance: -1.9984e-14 on 0 degrees of freedom
AIC: 101.15

Number of Fisher Scoring iterations: 3

```

شکل ۳.۵: نتایج استفاده از مدل اشباع شده با استفاده از دستور `glm`

```
loglin.out1 <- loglin(haireye.table, margin=list(1,2), param=TRUE)
```

و برای مدل اشباع شده در شکل (۵.۵) داریم:

```
loglin.out2 <- loglin(haireye.table, margin=list(c(1,2)), param=TRUE)
```

توجه کنید که برآوردها با استفاده از هر دو دستور `glm` و `loglin` یکسان هستند.  
حال فرض کنید متغیر جنس نیز به داده‌ها وارد شده باشد و جدول ۴.۵ را داشته باشیم.

#### جدول ۴.۵: داده‌های مربوط به رنگ مو و چشم تعدادی مرد

جنس (مرد) رنگ چشم					رنگ مو
سبز	فندقی	آبی	قهوه‌ای	قرمز	پور
۳	۱۰	۱۱	۳۲	مشکی	قهوه‌ای
۱۵	۲۵	۵۰	۳۸	قرمز	پور
۷	۷	۱۰	۱۰		
۸	۵	۳۰	۳		

جنس (زن)				
رنگ چشم				رنگ مو
سبز	آبی	فندقی	قهوه‌ای	
۲	۵	۹	۳۶	مشکی
۱۴	۲۹	۳۴	۸۱	قهوه‌ای
۷	۷	۷	۱۶	قرمز
۸	۵	۶۴	۴	پور

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیش‌بینی

```

$lrt
[1] 44.31537

$spearson
[1] 42.16325

$df
[1] 9

$margin
$margin[[1]]
[1] "hair"

$margin[[2]]
[1] "eye"

$param
$param$(Intercept)
[1] 2.577301

$param$hair
    Black      Blond      Brown       Red
-0.03274428 -0.22945457  0.79393429 -0.53173544

$param$eye
    Blue      Brown      Green      Hazel
 0.5199664  0.3236865 -0.5986465 -0.2450065

```

شکل ۴.۵: نتایج استفاده از مدل استقلال با استفاده از دستور loglin

```

$lrt
[1] 0

$spearson
[1] 0

$df
[1] 0

$margin
$margin[[1]]
[1] "hair" "eye"

$param
$param$(Intercept)
[1] 2.46419

$param$hair
    Black      Blond      Brown       Red
-0.1479831 -0.4170179  0.9049436 -0.3399426

$param$eye
    Blue      Brown      Green      Hazel
 0.5392350  0.1619397 -0.5061867 -0.1949880

$param$hair.eye
            eye
        hair      Blue      Brown      Green      Hazel
    Black -0.457546845  0.98758911 -0.7114082  0.18136592
    Blond  0.814790122 -1.11049964  0.5384559 -0.24274640
    Brown  0.003654229  0.10651271 -0.1548969  0.04472999
    Red   -0.360897506  0.01639782  0.3278492  0.01665049

```

شکل ۵.۵: نتایج استفاده از مدل اشباع‌شده با استفاده از دستور loglin

دستورهای زیر، داده‌های جدول ۴.۵ را وارد می‌کند و داده‌ها در قالب data frame چاپ می‌شوند.

```
hair <- factor(rep(rep(c("Black", "Brown", "Red", "Blond"),
c(4,4,4,4)),2))
eye <- factor(rep(c("Brown", "Blue", "Hazel", "Green"), 8))
sex <- factor(rep(c("Male", "Female"), c(16,16)))
freq <- c(32,11,10,3, 38,50,25,15, 10,10,7,7, 3,30,5,8,
36,9,5,2, 81,34,29,14, 16,7,7,7, 4,64,5,8)
haireye2.data <- data.frame(hair, eye, sex, freq)
haireye2.data
```

همانند قبل، دستور زیر، داده‌ها را به صورت جدول نمایش می‌دهد.

```
haireye2.table <- table(hair, eye, sex)
haireye2.table[cbind(hair, eye, sex)] <- freq
haireye2.table
```

دستور زیر، مدل استقلال را بر اساس دستور glm برازش می‌دهد که خروجی آن در شکل (۶.۵) آمده است:

```
glm.out3 <- glm(freq~hair+eye+sex, family=poisson,
data=haireye2.data)
summary(glm.out3)
```

چون مدل استقلال، برازش خوبی ندارد، مدل اشباع شده را به کار می‌بریم و تحلیل کیبیش را انجام می‌دهیم، که این کار در نرم‌افزار R با استفاده از دستورهای زیر صورت می‌گیرد. خروجی نتایج دستورهای زیر در شکل (۷.۵) نمایش داده شده است.

## فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشانی‌دهی

```

Call:
glm(formula = freq ~ hair + eye + sex, family = poisson, data = haireye2.data)

Deviance Residuals:
    Min      1Q   Median     3Q    Max 
-5.4109 -1.5996  0.2315  0.9084  6.3735 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 2.64265   0.05219 50.635 < 2e-16 ***
hair1       -0.17912   0.08246 -2.172  0.02984 *  
hair2       -0.01706   0.07814 -0.218  0.82718    
hair3        0.79474   0.06259 12.697 < 2e-16 ***
eye1        0.50670   0.06745  7.513 5.79e-14 ***
eye2        0.52969   0.06705  7.900 2.80e-15 ***
eye3        -0.70505   0.10018 -7.038 1.95e-12 ***
sex1         0.10853   0.04134  2.625  0.00866 ** 
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 489.59 on 31 degrees of freedom
Residual deviance: 175.79 on 24 degrees of freedom
AIC: 330.54

Number of Fisher Scoring iterations: 5

```

شکل ۶.۵: نتایج استفاده از مدل استقلال در داده‌ها، با در نظر گرفتن متغیر جنس و با استفاده از دستور `glm`

```

glm.out4 <- glm(freq~hair*eye*sex, family=poisson,
data=haireye2.data)

anova(glm.out4, test="Chisq")

```

اگر سطح معناداری را ۱٪ در نظر بگیریم، مدلی که همه‌ی اثرهای متقابل دوتایی را در بر دارد، بهترین مدل است. بنا بر این هیچ یک از دو متغیر، استقلال شرطی ندارد؛ ولی پیوند شرطی هر دو متغیر در سطوح متغیر دیگر، یکسان است (همگنی). دستور زیر، یک مدل غیرسلسله‌مراتبی را برازش می‌دهد که نتایج آن نیز در شکل (۸.۵) ارائه شده است.

```

glm.out5 <- glm(freq~sex+hair*eye, family=poisson,
data=haireye2.data)

anova(glm.out5, test="Chisq")

```

مقدار کیبس ۲۹/۳۵ با ۱۵ درجه‌ی آزادی نشان می‌دهد که این مدل باید رد شود. برای محاسبه‌ی  $P$ -مقدار از دستور زیر استفاده می‌کنیم:

۱.۵ دستورهای R برای برازش مدل‌های لگ‌خطی

۱۵۵

Analysis of Deviance Table						
	Model: poisson, link: log			Response: freq		
	Terms added sequentially (first to last)					
	Df	Deviance	Resid.	Df	Resid.	Dev P(> Chi )
NULL				31	489.59	
hair	3	165.59		28	324.00	1.138e-35
eye	3	141.27		25	182.73	2.010e-30
sex	1	6.93		24	175.79	0.01
hair:eye	9	146.44		15	29.35	4.806e-27
hair:sex	3	6.27		12	23.08	0.10
eye:sex	3	14.90		9	8.19	1.908e-03
hair:eye:sex	9	8.19		0	1.665e-14	0.52

شکل ۷.۵: نتایج استفاده از مدل اشباع شده در داده‌ها، با درنظر گرفتن متغیر جنس و با استفاده از دستور `glm`

Analysis of Deviance Table						
	Model: poisson, link: log			Response: freq		
	Terms added sequentially (first to last)					
	Df	Deviance	Resid.	Df	Resid.	P(> Chi )
NULL				31	489.59	
sex	1	6.93		30	482.66	0.01
hair	3	165.59		27	317.07	1.138e-35
eye	3	141.27		24	175.79	2.010e-30
hair:eye	9	146.44		15	29.35	4.806e-27

شکل ۸.۵: نتایج استفاده از مدل غیر سلسه مراتبی در داده‌ها، با درنظر گرفتن متغیر جنس و با استفاده از دستور `glm`

Analysis of Deviance Table						
Model: poisson, link: log						
Response: freq						
Terms added sequentially (first to last)						
	Df	Deviance	Resid.	Df	Resid.	Dev
NULL				31	489.59	P(> Chi )
sex	1	6.93		30	482.66	0.01
hair	3	165.59		27	317.07	1.138e-35
eye	3	141.27		24	175.79	2.010e-30
sex:eye	3	7.32		21	168.48	0.06
eye:hair	9	146.44		12	22.03	4.806e-27

شکل ۹.۵: نتایج استفاده از مدل غیر سلسه مراتبی دیگر در داده‌ها، با در نظر گرفتن متغیر جنس و با استفاده از دستور `glm`

```
1 - pchisq(29.35, 15)
```

```
[1] 0.01449365
```

دستورهای مورد نیاز برای تحلیل یک مدل غیر سلسه مراتبی دیگر به صورت زیر است، که نتایج آن نیز در شکل (۹.۵) ارائه شده است.

```
glm.out6 <- glm(freq~sex+eye:sex+hair*eye, family=poisson,
data=haireye2.data)
anova(glm.out6, test="Chisq")
```

اکنون با استفاده از نتایج دستور زیر در می‌یابیم که اثرهای متقابل جنس و رنگ چشم با  $P$ -مقداری برابر با  $6/00\%$  معنادار است. کیبیش این مدل،  $3/02$  با  $22/03$  درجه‌ی آزادی است. بنا بر این، این مدل در مقابل مدل اشباع‌شده رد می‌شود.

1 - `pchisq(22.03, 12)`

[1] 0.03718496

## ۹.۵ تمرین‌ها

۱- جدول سه‌طرفه‌ای با  $I \times J \times K$  خانه را در نظر بگیرید. مدل لگ‌خطی زیر را در نظر بگیرید:

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}.$$

آ) محدودیت‌های لازم بر روی پارامترها را ذکر کنید.

ب) بر حسب  $\mu_{ijk}$  عبارتی برای  $\lambda_{ij}^{XY}$  بیابید.

۲- در یک جدول  $2 \times 2 \times 2$ ، نشان دهید که اثرهای متقابل سه‌طرفه‌ی یک مدل لگ‌خطی اشعاع‌شده، لگاریتم اندازه‌ی بارتلت (۱۹۳۵) زیر است:

$$\frac{\mu_{111}\mu_{122}\mu_{212}\mu_{221}}{\mu_{112}\mu_{121}\mu_{211}\mu_{222}}$$

نشان دهید که این اندازه‌ی نسبت، نسبت‌های بخت در دو جدول  $2 \times 2$  است.

۳- مدل لگ‌خطی زیر برای یک جدول  $J \times I$  با محدودیت‌های سطر آخر صفر را در نظر گیرید.

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

نشان دهید که  $\lambda_{ij}^{XY}$  لگاریتم نسبت بخت‌های یک جدول  $2 \times 2$  است. این جدول را به عنوان قسمتی از جدول بزرگ‌تر  $J \times I$  مشخص کنید.  
در یک جدول سه‌بعدی، برای مدل لگ‌خطی  $(XZ, YZ)$ ، آماره‌های بسنده‌ی کمینه، معادلات درست‌نمایی، مقادیر برازانده‌شده و درجه‌ی آزادی مانده‌ها را بیابید و الگوریتم برازش متناسب تکراری (IPF) را به کار ببرید.

فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشاپنده‌ی

۴- جدول چهارطرفه‌ای را که متغیرهای عاملی آن  $W$  و  $X$  و  $Y$  و  $Z$  هستند در نظر بگیرید.

(آ) برای مدل  $(WXZ, WYZ)$  نشان دهید که  $X$  و  $Y$  به شرط  $W$  و  $Z$  مستقل‌اند. آیا نوع دیگری از استقلال شرطی در این مدل وجود دارد؟

(ب) برای چه مدل‌هایی که اثر متقابل سه‌تایی را در بر ندارند،  $X$  و  $Y$  به شرط  $W$  و  $Z$  مستقل‌اند؟

۵- اشفورد و سودن (۱۹۷۰) داده‌های جدول (۵.۵) را گردآوری کرده‌اند. شمارش‌های این جدول، افراد حفار معدن را نشان می‌دهد که سن آن‌ها از ۲۰ تا ۶۴ سال است و سیگاری هستند ولی علائم بیماری قلبی ندارند. در این جدول، افراد بر حسب گروه سنی ( $X$ )، آیا نفس‌نفس می‌زنند یا نه ( $Y$ )، و آیا سینه‌ی آن‌ها خس‌خس می‌کند یا نه ( $Z$ ) ردیابنده‌ی شده‌اند.

جدول ۵.۵: داده‌های مربوط به ویژگی‌های افراد حفار معدن

گروه سنی	نفس نفس می‌زنند	نفس نفس نمی‌زنند			کل
		خس خس دارند	خس خس ندارند	خس خس دارند	
۲۴-۲۰	۹	۷	۹۵	۱۸۴۱	۱۹۵۲
۲۹-۲۵	۲۳	۹	۱۰۵	۱۶۵۴	۱۷۹۱
۲۳۴-۳۰	۵۴	۱۹	۱۷۷	۱۸۶۳	۲۱۱۳
۳۹-۴۵	۱۲۱	۴۸	۲۵۷	۲۲۵۷	۲۷۸۳
۴۴-۴۰	۱۶۹	۵۴	۲۲۳	۱۷۷۸	۲۲۷۴
۴۹-۴۵	۲۶۹	۸۸	۳۲۴	۱۷۱۲	۲۳۹۳
۵۴-۵۰	۴۰۴	۱۱۷	۲۴۵	۱۳۲۴	۲۰۹۰
۵۹-۵۵	۴۰۶	۱۰۲	۲۲۵	۹۶۷	۱۷۵۰
۶۴-۶۰	۳۷۲	۱۰۶	۱۳۲	۵۲۶	۱۱۳۶
کل	۱۸۲۷	۶۰۰	۱۸۳۳	۱۴۰۲۲	۱۸۲۸۲

(آ) مدل لگ خطی که سن را به عنوان عامل در نظر می‌گیرد و اثرهای متقابل دودویی را در بر دارد، برازش دهید.

(ب) برای هر گروه سنی، لگاریتم نسبت بخت‌ها برای پیوند بین  $Z$  و  $Y$  را بیابید. سپس نموداری از این مقادیر در مقادیر  $x$  رسم کنید که در آن  $x$  نقطه‌ی وسط گروه سنی  $\bar{x}$  است.

(پ) مدل لگ خطی زیر را در نظر بگیرید و آن را برازش دهید. آیا می‌توانید مدل بهتری

بیابید؟

$$\log \mu_{ijk} = \lambda + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$$

۶- یک مدل چهار طرفه با متغیرهای  $R$  و  $X$  و  $Y$  و  $Z$  را در نظر بگیرید. فرض کنید که می‌توانید  $R$  را به عنوان یک متغیر پاسخ در نظر بگیرید. در هر حالت زیر، مدل لگ خطی با پاسخ پواسون مریبوط را بنویسید.

(۱) از  $X$  به طور شرطی مستقل است.

(۲) فقط  $X$  و  $Z$  روی  $R$  تأثیر دارند، اما تأثیرشان اثر متقابل ندارد.

(۳) اثر متقابل سه‌تایی وجود ندارد.

۷- نشان دهید که کلی ترین مدل لگ خطی  $K$ -بعدی (با  $K$  متغیر)،  
 $2^K$  جمله دارد.

۸- در مدل لگ خطی سلسه مراتبی ( $XY, XZ, XW$ ) با چهار متغیر  $X$  و  $Y$  و  $Z$  و  $W$ ،

(آ) آماره‌های بسته‌ده را بیابید.

(ب) آیا صورت بسته‌ای برای مقادیر برازنده شده  $\mu_{ijkl}$  وجود دارد؟ اگر جواب شما بلی است، آن را بیابید.

(پ) نیکویی برازش این مدل را در مقابل یک مدل اشباع شده، با یافتن آماره‌ی آزمون مناسب و توزیع این آماره تحت این مدل، تشریح کنید.

۹- جدول زیر را در نظر بگیرید.

$Y$		$X$
۲	۱	
۱۰۰	۱۲۶	۱
۶۱	۳۵	۲
۶۸۸	۹۰۸	۳

برای مدل  $\log \mu_{ij} = \lambda + \lambda_i^X$  برآورد  $\mu_{ij}$  را بیابید و نیکویی برازش مدل را آزمون کنید.

فصل ۵. مدل‌های لگ خطی برای جدول‌های پیشانی‌دهی

۱۰ - جدول زیر را در نظر بگیرید:

$Y$			
	$X$		$Z$
۲	۱	۱	۱
۱۰۰	۱۲۶		
۶۱	۳۵	۲	
۹۰۸	۶۸۸	۱	۲
۴۹۷	۸۰۷	۲	

برای مدل  $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$  برآورد  $\mu_{ijk}$  را تحت این مدل بیابید و نیکویی برازش مدل را آزمون کنید. اگر  $Y$  متغیر پاسخ باشد، مدل لوجیت متناظر مدل بالا را بیابید.

۱۱ - در شکل‌های ۸.۵ و ۹.۵ به ترتیب، چه مدل‌های غیر سلسه مراتبی‌ای در نظر گرفته شده‌اند؟ آیا این مدل‌ها مناسب‌اند؟