

فصل ۲

مطالعات مقطعی: داده‌های دارای

پاسخ دودویی

یکی از انواع مطالعاتی که در این کتاب مورد توجه است، مطالعات مقطعی با پاسخ‌های دودویی است. در این فصل، نخست نمادگذاری‌های مورد استفاده برای مطالعات مقطعی بیان می‌شود و به تشریح این نوع داده‌ها پرداخته می‌شود. سپس توزیع‌های مورد استفاده برای پاسخ‌های دودویی بررسی می‌شود. هنگامی که با پاسخ‌های دودویی مواجه هستیم مدل‌ها و روش‌های مختلفی مورد استفاده قرار می‌گیرند که گاهی استفاده از آن‌ها پیچیده‌تر از هنگامی است که متغیر پاسخ پیوسته است. مدل‌هایی که عموماً از آن‌ها استفاده می‌شود، مدل‌های رگرسیون لوژستیک، پروبیت، لگ-لگ-لگ مکمل‌اند.

توزیع‌های مورد استفاده برای پاسخ‌های دودویی و مدل‌های مختلف مورد استفاده برای تحلیل این گونه داده‌ها در بخش‌های بعدی معرفی می‌شوند. سپس روش‌های محاسباتی برای تحلیل این نوع داده‌ها، و کاربرد این روش‌ها در مثال‌های کاربردی، مورد بحث قرار می‌گیرند. معیارهای نیکویی برازش در مطالعات دارای پاسخ دودویی و دستورهای مورد استفاده در نرم‌افزار R برای تحلیل داده‌های دودویی از مباحث پایانی این فصل از کتاب می‌باشند.

۱.۲ مفاهيم و نمادگذاري

همان طور که در فصل ۱ اشاره شد، به آن دسته از مطالعات که در آن‌ها فقط یک متغير پاسخ در مقطعي معلوم از زمان مورد بررسی است و می‌توان آن را به وسیله‌ی یک بردار $x = (x_1, \dots, x_k)'$ از متغيرهای کمکی یا تبیینی تعبیر کرد، مطالعات مقطعي می‌گویند. در این مطالعات، متغير پاسخ می‌تواند پیوسته یا گسسته، شمارشی یا دودویی باشد و متغيرهای کمکی می‌توانند پیوسته یا ردده‌بندی شده از نوع ترتیبی یا اسمی باشند. در این فصل، فرض خواهد شد که متغير پاسخ، \mathcal{Y} ، فقط دو مقدار اختیار می‌کند و بدون این که خللی به کلیت موضوع وارد شود، فرض خواهد شد که مقادیر صفر و یک را می‌گیرد. شایان ذکر است که مطالعاتی نظیر رگرسیون، تحلیل واریانس و تحلیل کوواریانس برای پاسخ‌های دودویی نیز در این فصل، مورد مطالعه قرار می‌گیرند.

به عنوان مثالی از مطالعات مقطعي، فرض کنید متغير پاسخ، بهبودی یا عدم بهبودی بیمار در طول یک دوره‌ی درمان است که بهبودی بیمار را با «(یک)» و عدم بهبودی را با «(صفراً)» نشان می‌دهیم. هدف از چنین مطالعه‌ای می‌تواند بررسی تأثیر سن و نوع درمان بر بهبودی بیمار باشد؛ سن، یک متغير پیوسته است و نوع درمان، یک متغير گسسته. سن و نوع درمان، متغيرهای تبیینی یا کمکی در این مطالعه‌اند.

۲.۲ توزيع مناسب برای داده‌های مقطعي دارای پاسخ دودویی

محاسبه‌ی احتمال‌ها در بعضی موارد، بسیار مشکل است و یکی از روش‌ها برای ساده کردن محاسبه‌ی احتمال‌ها استفاده از توزيع‌های آماری برای برخی از متغيرهای تصادفي است که دارای شرایط خاص و معینی هستند. یکی از این توزيع‌ها توزيع برنولی است که برای متغيرهای دودویی به کار می‌رود.

هنگامی که متغيری فقط دو حالت موفقیت ($1 = Y$) و شکست ($0 = Y$) را اختیار می‌کند، گفته می‌شود این متغير (Y) دارای توزيع برنولی با احتمال موفقیت p

است. در این حالت، تابع جرم احتمال متغیر تصادفی Y به صورت زیر

است:

$$\Pr(Y = y) = p^y(1 - p)^{1-y}, \quad y = 0, 1$$

در این صورت به اختصار می‌نویسند $Y \sim \text{Ber}(p)$. همان‌گونه که در کتاب‌های آمار

ریاضی دیده‌ایم:

$$E(Y) = p, \quad \text{var}(Y) = p(1 - p).$$

به عنوان مثال، یک سکه‌ی اریب با احتمال شیر آمدن $\frac{1}{4}$ را در نظر گیرید. اگر سکه یک بار پرتاب شود، بسته به آن که خط یا شیر ظاهر شود، متغیر تصادفی Y به ترتیب، مقادیر ۰ و ۱ را اختیار می‌کند. تابع جرم احتمال متغیر تصادفی Y برابر است با

$$\Pr(Y = y) = (0/4)^y(1 - 0/4)^{1-y}, \quad y = 0, 1$$

که از آن $0/4 = 0/6 = \Pr(Y = 0) = \Pr(Y = 1) = 0/6 = 0/24 = 0/24$ و واریانس آن برابر با ۰ است. (برای دیگر خواص توزیع برنولی به پارسیان، ۱۳۷۸، صص. ۱۵–۱۶ رجوع کنید).

حالتی از تعمیم توزیع برنولی، توزیع دوجمله‌ای است. اگر متغیرهای Y_1, Y_2, \dots, Y_n متغیرهای تصادفی مستقل از توزیع برنولی با احتمال موفقیت p باشند، مجموع این متغیرها $(Y = \sum_{i=1}^n Y_i)$ متغیری تصادفی از توزیع دوجمله‌ای با پارامترهای n و p است؛ یعنی تابع جرم احتمال این متغیر به صورت زیر است:

$$\Pr(Y = y) = \binom{n}{y} p^y(1 - p)^{n-y}, \quad y = 0, 1, \dots, n$$

در این صورت می‌نویسند $Y \sim \text{Bin}(n, p)$.

متغیر تصادفی Y با توزیع دوجمله‌ای، دارای میانگین np و واریانس $(p - np)(1 - np)$ است. (برای مطالعه‌ی بیشتر درباره‌ی ویژگی‌های توزیع دوجمله‌ای به پارسیان، ۱۳۷۸، ص. ۱۶ مراجعه کنید).

در مبحث آمار ریاضی، فرض می‌شود که احتمال موفقیت در توزیع برنولی یا توزیع دوجمله‌ای، p ، نامعلوم است و برای برآورده کردن آن، نمونه‌ی تصادفی Y_1, \dots, Y_n استخراج

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

می‌شود که مستقل و از توزیع یکسان برآورده‌ای است (به این معنا که «از یک نمونه به نمونه‌ی دیگر یا از یک فرد به فرد دیگر، تغییر نمی‌کند»). در ادامه، فرض می‌کنیم که نمونه‌های ما مستقل‌اند، ولی احتمال موفقیت p ، از طریق برداری از متغیرهای تبیینی، وابسته به خصوصیات فرد است.

۳.۲ مدل‌های مختلف برای تحلیل داده‌های مقطعی دارای پاسخ دودویی

۱.۳.۲ مدل احتمالاتی خطی

در این بخش، مدل احتمالاتی خطی معرفی می‌گردد و نشان داده می‌شود که دارای اشکالاتی اساسی است. لذا در بخش‌های بعدی مدل‌های دیگری برای تحلیل داده‌های دارای پاسخ دودویی ارائه خواهند شد.

مدل احتمالاتی خطی، مدل رگرسیونی شامل یک متغیر وابسته‌ی دودویی است. ساختار مدل از حالتی می‌آید که y_i یک متغیر پیوسته است. در حالتی که y_i پیوسته است، مدل احتمالاتی خطی عبارت است از

$$y_i = x'_i \beta + \varepsilon_i,$$

که در آن x_i یک بردار از مقادیر متغیرهای تبیینی ثبت‌شده برای نامین آزمودنی است، β یک بردار از پارامترها است و ε جمله‌ی خطأ با میانگین صفر است. به عنوان مثال، اگر در مدلی فقط یک متغیر مستقل داشته باشیم، داریم:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

$$\text{در این صورت، } E(Y_i|x_i) = \alpha + \beta x_i.$$

برای درک مدل احتمالاتی خطی، باید مفهومی از $E(Y|x)$ را در نظر بگیریم. هنگامی که y_i دودویی هستند، امید ریاضی غیرشرطی Y برابر با احتمال رخ دادن پیشامد است؛ یعنی

$$E(Y_i) = 1 \times \Pr(Y_i = 1) + 0 \times \Pr(Y_i = 0) = \Pr(Y_i = 1).$$

در حالت کلی برای هر مدل رگرسیونی، امید ریاضی شرطی به صورت زیر است:

$$E(Y_i|x_i) = 1 \times \Pr(Y_i = 1|x_i) + 0 \times \Pr(Y_i = 0|x_i) = \Pr(Y_i = 1|x_i).$$

بنا بر این با داشتن مقدار x_i ، میانگین Y_i برابر است با احتمال $1 = Y_i$. این به ما اجازه می‌دهد که مدل احتمالاتی خطی برای پاسخ دودویی را به صورت زیر تعریف کنیم:

$$\Pr(Y_i = 1|x_i) = x'_i \beta. \quad (1.2)$$

دودویی بودن متغیر پاسخ، تأثیری بر تفسیر پارامترها ندارد و پارامترها همانند رگرسیون معمولی تفسیر می‌شوند؛ به این معنا که به ازای یک واحد افزایش در x_{ik} (یعنی k امین متغیر تبیینی فرد i)، با فرض ثابت نگه داشتن متغیرهای دیگر، احتمال موفقیت به اندازه β_k (یعنی k امین ضریب رگرسیونی متناظر با x_{ik}) تغییر می‌کند.

اشکال‌های مدل احتمالاتی خطی

۱ - ناهمگنی واریانس‌ها

اگر یک متغیر تصادفی دودویی دارای میانگین μ باشد، واریانس آن برابر با $(\mu - 1)\mu$ خواهد بود. چون امید ریاضی Y به شرط x برابر با $x'\beta$ است، واریانس Y به شرط x به صورت زیر است:

$$\text{var}(Y|x) = \Pr(Y = 1|x)[1 - \Pr(Y = 1|x)].$$

در نتیجه:

$$\text{var}(Y|x) = x'\beta(1 - x'\beta),$$

که نشان می‌دهد واریانس خطای وابسته به x است و از یک فرد به فرد دیگر تغییر می‌کند. بنا بر این براوردگر کمترین توانهای دوم معمولی β کارا نیست و خطاهای استاندارد این براوردگرها، اریب می‌باشند (سیرل، ۱۹۷۱).

گولدبرگ (۱۹۶۴) پیشنهاد کرد که مدل‌های احتمالاتی خطی برای تصحیح ناهمگنی واریانس از براوردگر دومرحله‌ای استفاده کنند: در اولین مرحله، همان براوردگر کمترین توانهای دوم معمولی، $\hat{\beta}$ ، و در دومین مرحله، مدل به وسیله‌ی کمترین توانهای دوم

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

تعمیم‌یافته با استفاده از $\widehat{\text{var}}(\hat{\varepsilon}) = (\hat{y} - 1)(\hat{y})$ برای تصحیح ناهمگنی واریانس‌ها براورد شود. این راه با آن که کارایی را افزایش می‌دهد، با مشکل دیگری مواجه است. این شیوه برای $1 < \hat{y}$ یا $\hat{y} < 0$ ، واریانس را منفی براورد می‌کند.

۲- نرمال بودن

برای یک مقدار خاص x ، مانند x_* ، خطاهای مقادیر $E(Y|x_* - 1) = \mu$ یا $E(Y|x_* - 0) = \sigma$ را اختیار می‌کنند. واضح است که خطاهای نمی‌توانند به صورت نرمال توزیع شده باشند.

۳- براوردهای غیرحساس

در مدل احتمالاتی خطی، احتمال موفقیت پاسخ دودویی به ازای مقادیر به اندازه‌ی کافی کوچک یا بزرگ x ، به صورت منفی یا بزرگ‌تر از یک، براورد می‌شود.

معمولًاً اثر متغیرهای تبیینی بر احتمال موفقیت، صورت خطی ندارد و یک صورت غیر خطی مناسب‌تر است. در بیش‌تر کاربردها اثر یک متغیر تبیینی پیوسته‌ی خاص بر احتمال موفقیت، S-شکل می‌باشد. قبل از پرداختن به این موضوع، اجازه دهید تا مدل خطی تعیین‌یافته را مورد بررسی قرار دهیم، که مدل‌های دارای پاسخ دودویی، حالت خاص آن هستند.

۲.۳.۲ مدل‌های خطی تعیین‌یافته

در این زیربخش به بررسی مدل‌های خطی تعیین‌یافته خواهیم پرداخت و مؤلفه‌های آن را شرح خواهیم داد. سپس آن را برای حالت پاسخ دودویی مورد بحث قرار خواهیم داد، که شامل مدل رگرسیون لوزیستیک، پروبیت، لگ-لگ-لگ مکمل است.

۱.۲.۳.۲ تعریف مدل‌های خطی تعیین‌یافته

مدل‌های خطی تعیین‌یافته که از این به بعد آن‌ها را با GLM^۱ نشان می‌دهیم، به وسیله‌ی سه مؤلفه‌ی زیر مشخص می‌شوند:

۱- مؤلفه‌ی تصادفی،

۲- مؤلفه‌ی سیستماتیک،

۳- تابع ربط.

۱ - **مؤلفه‌ی تصادفی:** مؤلفه‌ی تصادفی، توزیع احتمالی متغیر پاسخ را مشخص می‌کند. این مؤلفه شامل مشاهدات مستقل y_1, \dots, y_n از متغیر تصادفی Y با توزیعی از یک خانواده‌ی نمایی طبیعی است. هر یک از این مشاهدات، دارای تابع چگالی زیر است:

$$f(y_i, \theta_i) = a(\theta_i)b(y_i) \exp\{y_i Q(\theta_i)\}, \quad (2.2)$$

که $Q(\theta)$ را پارامتر طبیعی می‌نامند. به عنوان مثال، توزیع‌های برنولی، پواسون، چندجمله‌ای و نرمال که بسیار مورد استفاده‌اند، مثال‌هایی از اعضای این خانواده می‌باشند.

۲ - **مؤلفه‌ی سیستماتیک:** مؤلفه‌ی سیستماتیک شامل یک بردار η است که ترکیبی خطی از مجموعه‌ی متغیرهای تبیینی می‌باشد؛ یعنی

$$\eta = x' \beta,$$

که در آن x یک بردار از متغیرهای تبیینی است و β یک بردار از پارامترهای مدل است. η نیز یک برآورده‌گر خطی است که پیشگوی خطی نامیده می‌شود.

۳ - **تابع ربط:** ارتباط بین مؤلفه‌ی سیستماتیک ($\eta = x' \beta$) و مقادیر مورد انتظار مؤلفه‌های تصادفی را توصیف می‌کند. تابع ربط، تابعی است مشتق‌پذیر و یکنوا که میانگین پاسخ را به مؤلفه‌ی سیستماتیک پیوند می‌دهد. گیریم $g(\mu_i) = E(Y_i)$ ، که در آن g یک تابع مشتق‌پذیر یکنوا است. در نتیجه:

$$g(\mu_i) = x_i \beta = \sum_j x_{ij} \beta_j, \quad i = 1, \dots, N$$

به عنوان مثال، اگر $\mu_i = g(\mu_i)$ باشد، داریم $\sum_j x_{ij} \beta_j = \mu_i$ ، که به آن، تابع ربط همانی می‌گویند. همچنین اگر داشته باشیم $Q(\theta_i) = g(\mu_i)$ ، آن‌گاه $\sum_j x_{ij} \beta_j = Q(\theta_i)$. به این تابع ربط، ربط کانونی می‌گویند. جدول ۱.۲ انواع مختلف GLM‌ها را بیان می‌کند. در این جدول، به عنوان مثال، در مدل‌های تحلیل واریانس، مؤلفه‌ی سیستماتیک می‌تواند شامل متغیرهای تصادفی رسته‌ای باشد که با تابع ربط همانی به میانگین پاسخ مرتبط‌اند و مؤلفه‌ی تصادفی این مدل دارای توزیع نرمال است.

برای بسیاری از GLM‌ها (به خصوص مدل‌هایی که در این بخش معرفی خواهیم کرد)

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

لگاریتم درست‌نمایی اکیداً مفعر است و براوردگرهای ماکسیمم درست‌نمایی پارامترهای مدل، تحت شرایط نظم، به‌طور یکتا وجود دارند (ودربرن، ۱۹۷۴).

جدول ۱.۲: انواع مختلف مدل‌های خطی تعیین‌یافته (GLM)

مدل	مؤلفه‌ی سیستماتیک	تابع ربط کانونی	مؤلفه‌ی تصادفی
رگرسیون	پیوسته	همانی	نرمال
تحلیل واریانس	رستمای	همانی	نرمال
تحلیل کوواریانس	مختلط	همانی	نرمال
رگرسیون لوژستیک	مختلط	لوجیت	برنولی
لگ خطی	مختلط	لگاریتم	پواسون
پاسخ اسمی با چند سطح	مختلط	لوجیت تعیین‌یافته	چندجمله‌ای

گیریم $(y; \mu) \ell$ نشان‌دهنده‌ی لگاریتم درست‌نمایی بر حسب $(\mu_1, \dots, \mu_n) = \mu$ برای یک بردار داده‌های مشاهده شده، $(y_1, \dots, y_n) = y$ باشد. فرض کنید $(\hat{\mu}_M; y) \ell$ لگاریتم درست‌نمایی ماکسیمم تحت فرض مناسب بودن مدل M_1 است و لگاریتم درست‌نمایی ماکسیمم برای مدلی که دارای تعداد پارامترهایی برابر با تعداد مشاهدات است (که مدل اشباع شده مثالی از این حالت است) برابر با $\ell(y; y)$ باشد، که با جایگذاری $y_i = \hat{\mu}_i$ در $(y; \mu) \ell$ به دست می‌آید. کیبیش مدل M_1 به صورت زیر تعریف می‌شود:

$$D(M_1) = 2[\ell(y; y) - \ell(\hat{\mu}_M; y)], \quad (3.2)$$

که این کیبیش، هنگامی که مدل M_1 صادق باشد، یک توزیع مجانبی χ^2 دارد. درجه‌ی آزادی این توزیع، برابر با تفاضل تعداد مشاهدات (n) و تعداد پارامترهای موجود در مدل M_1 است. کیبیش می‌تواند به عنوان معیاری برای نیکویی برآش به کار رود. در پاسخ‌های دودویی، برای آماره‌ی کیبیش، فرمولی در زیربخش ۲.۵.۲ ارائه شده است. برای مقایسه‌ی مدل M_0 در مقابل مدل M_1 که پارامترهای بیشتری نسبت به M_0 دارد، آماره‌ی نسبت درست‌نمایی $[\ell(\hat{\mu}_M; y) - \ell(\hat{\mu}_{M_0}; y)] / 2$ است که می‌تواند با

$$\begin{aligned} LR(M_0 | M_1) &= 2[\ell(y; y) - \ell(\hat{\mu}_{M_0}; y)] - 2[\ell(y; y) - \ell(\hat{\mu}_{M_1}; y)] \\ &= D(M_0) - D(M_1) \end{aligned}$$

بیان شود. بنا بر این آماره‌ی نسبت درست‌نمایی در این حالت، اختلاف کیبیش‌های دو مدل است، که اگر این اختلاف کوچک باشد، به این معنا است که دو مدل به خوبی یکدیگر به

داده‌ها برازش داده شده‌اند. تحت مدل M_0 ، این اختلاف برای نمونه‌های بزرگ، دارای توزیع خی دو با درجه‌ی آزادی برابر با اختلاف پارامترهای دو مدل M_0 و M_1 است.

۲.۲.۳. ۲ GLM‌ها برای پاسخ‌های دودویی

بسیاری از متغیرهای پاسخ (Y) شامل دو برامد ممکن‌اند که آن‌ها را با صفر و یک نشان می‌دهیم. در واقع، Y یک متغیر تصادفی برنولی با $\Pr(Y = 1) = \pi$ می‌باشد. برای متغیرهای تصادفی دودویی با توزیع برنولی، داریم:

$$\begin{aligned} f(y_i; \pi_i) &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= (1 - \pi_i) \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} \\ &= (1 - \pi_i) \exp \left\{ y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) \right\}, \quad y_i = 0, 1 \end{aligned}$$

که نشان می‌دهد f عضو خانواده‌ی نمایی طبیعی با پارامتر طبیعی لوچیت π_i (لگاریتم بخت برای پاسخ ۱) است؛ یعنی:

$$\text{logit}(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right).$$

گشتاورهای اول و دوم Y_i به صورت زیر می‌باشند:

$$E(Y_i) = 1 \times \Pr(Y_i = 1) + 0 \times \Pr(Y_i = 0) = \Pr(Y_i = 1) = \pi_i$$

$$E(Y_i^2) = 1 \times \Pr(Y_i = 1) + 0 \times \Pr(Y_i = 0) = \Pr(Y_i = 1) = \pi_i$$

$$\text{var}(Y_i) = \pi_i(1 - \pi_i)$$

پس GLM برای پاسخ‌های دودویی، رابطه‌ی غیر خطی بین x_i ها و $\pi_i = \pi(x_i)$ را بیان می‌کند، که اغلب یک رابطه‌ی S-شکل بین احتمال و مقادیر x_i برقرار می‌کند (π را به صورت $\pi(x_i)$ بیان می‌کنیم تا بر وابستگی π_i به x_i ها تأکید کرده باشیم).

آ) رگرسیون لوژیستیک

در این نوع GLM،تابع ربط، همان پارامتر طبیعی توزیع، یعنی لوچیت ($\pi(x)$) است؛ یعنی:

$$\text{logit}[\pi(x)] = \log \left[\frac{\pi(x)}{1 - \pi(x)} \right] = x' \beta, \quad (4.2)$$

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

$$\pi(x) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}, \quad \text{یا به طور معادل،}$$

که به این تابع، رگرسیون لوژستیک می‌گویند.

گیریم یک متغیر توضیحی x داریم. در این صورت، $\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$. هنگامی که $\alpha \rightarrow -\infty$ و $\beta \rightarrow 0$ ، $\pi(x) \downarrow 0$ و به ازای $\beta > 0$ ، $\pi(x) \uparrow 1$. هنگامی که $\beta = 0$ باشد، منحنی به صورت یک خط راست افقی در می‌آید و پاسخ دودویی از x مستقل می‌باشد.

برای این مدل، بخت موفقیت به صورت زیر است:

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\alpha + \beta x) = e^\alpha (e^\beta)^x.$$

این فرمول، یک تفسیر اساسی برای β مهیا می‌کند. به ازای یک واحد افزایش x ، در بخت، افزایش ضربی e^β حاصل می‌شود.

برای تعیین یک شکل مناسب برای مؤلفه‌ی سیستماتیک در یک مدل رگرسیون لوژستیک، ترسیم نمودار پراکنش لوجیت‌های نمونه‌ای در برابر x ، مفید است. فرض کنید n_i تعداد مشاهدات در i -امین مجموعه‌ی x_i ها باشد و y_i تعداد یک‌ها در i -امین مجموعه. n_i لوجیت نمونه‌ای برابر است با $\text{logit}_i = \log\left(\frac{y_i}{n_i - y_i}\right)$ ، که این به ازای مقادیر $y_i = n_i$ و $y_i = 0$ جوابی ندارد. برای همین منظور، از لوجیت تجربی استفاده می‌کنیم، که به صورت زیر تعریف می‌شود:

$$\text{logit}_i = \log\left(\frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}\right). \quad (5.2)$$

هنگامی که x_i پیوسته است، رده‌بندی داده‌ها قبل از محاسبه‌ی لوجیت‌های تجربی، ضروری می‌باشد.

مدل‌های احتمال ناخطری دیگر به صورت $F(x'\beta) = F(x'\beta)$ می‌باشند که در آن $F(\cdot)$ در بازه‌ی $[0, 1]$ تغییر می‌کند. تابع‌های توزیع تجمعی، مثال‌هایی از توابع $F(\cdot)$ هستند که دارای این خاصیت می‌باشند. در واقع آن‌ها GLM‌هایی با تابع ربط F^{-1} می‌باشند. تعدادی از آن‌ها در زیر آمده‌اند.

ب) مدل پریویت

اگر در مدل $\pi(x) = F(x'\beta)$ به جای F از Φ (تابع توزیع تجمعی نرمال استاندارد) استفاده شود، مدل پریویت حاصل می‌شود؛ یعنی:

$$\pi(x) = \Phi(x'\beta). \quad (7.2)$$

مدل‌های پریویت و مدل لوجیت، متقارن هستند. به همین دلیل، سرعت میل کردن $\pi(x)$ به صفر برابر با سرعت میل کردن آن به ۱ در هر یک از مدل‌ها است. مدل‌های پریویت و مدل‌های لوجیت، هنگامی که $\pi(x)$ سرعتی متفاوت در نزدیک شدن به صفر و یک دارد، نامناسب می‌باشند و مدل‌هایی را باید در نظر گرفت که نامتقارن هستند.

پ) مدل لگ-لگ مکمل

در مدل لگ-لگ مکمل، تابع ربط به صورت $\log\{-\log[1 - \pi(x)]\}$ است؛ یعنی مدل به صورت زیر است:

$$\log\{-\log[1 - \pi(x)]\} = x'\beta$$

یا به طور معادل،

$$\pi(x) = 1 - \exp[-\exp(x'\beta)]. \quad (7.2)$$

اگر فرض کنیم که در مدل، فقط یک متغیر تبیینی داریم، مدل به صورت زیر خواهد بود:

$$\log\{-\log[1 - \pi(x)]\} = \alpha + \beta x.$$

در تفسیر این مدل، اگر دو مقدار x_1 و x_2 را برای x در نظر بگیریم، داریم:

$$\log\{-\log[1 - \pi(x_2)]\} - \log\{-\log[1 - \pi(x_1)]\} = \beta(x_2 - x_1),$$

به طوری که

$$\frac{\log[1 - \pi(x_2)]}{\log[1 - \pi(x_1)]} = \exp[\beta(x_2 - x_1)],$$

$$\cdot 1 - \pi(x_2) = (1 - \pi(x_1))^{\exp[\beta(x_2 - x_1)]}.$$

احتمال شکست در x_2 ، برابر است با احتمال شکست در x_1 به توان e^β به ازای هر یک

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

واحد اختلاف در فاصله‌ی بین x_1 و x_2 . در مدل لگ-لگ مکمل، سرعت رساندن $\pi(x)$ به ۱، از سرعت رساندن $\pi(x)$ به صفر بیشتر است. هنگامی که $0 < \beta$ ، مدل برای $\pi(x)$ نزولی است و به ازای $0 > \beta$ ، مدل برای $\pi(x)$ صعودی خواهد بود. توجه داشته باشید که اگر در $F(x'\beta) = F(\pi(x))$ تابع F را تابع توزیع گامبل یا توزیع نقاط فرین در نظر بگیریم، مدل لگ-لگ مکمل به دست می‌آید.

ت) مدل لگ-لگ
مدل لگ-لگ به صورت زیر است:

$$\pi(x) = \exp\{-\exp(x'\beta)\} \quad (8.2)$$

که در صورت وجود تنها یک متغیر تبیینی، به صورت $\pi(x) = \exp\{-\exp(\alpha + \beta x)\}$ خواهد بود. در مدل لگ-لگ، بر عکس مدل لگ-لگ مکمل، سرعت رساندن $\pi(x)$ به صفر، بیشتر از سرعت رساندن به یک است. همچنین هنگامی که $0 > \beta$ باشد، منحنی $\pi(x)$ نزولی خواهد بود و وقتی $0 < \beta$ باشد، منحنی آن صعودی خواهد بود. می‌توان مدل

$$\pi(x) = \exp\{-\exp(\alpha + \beta x)\}$$

یا به طور معادل،

$$\log\{-\log[\pi(x)]\} = \alpha + \beta x$$

را یک GLM با تابع ربط $\log\{-\log[\pi(x)]\}$ در نظر گرفت.

هنگامی که مدل لگ-لگ مکمل برای احتمال موفقیت، صادق باشد، مدل لگ-لگ برای احتمال شکست، صادق خواهد بود.

مدل لگ-لگ را می‌توان به صورت $F(x'\beta) = F(\pi(x))$ با حالت خاصی از تابع توزیع تجمعی گامبل فرض کرد. می‌دانیم که تابع توزیع تجمعی گامبل به صورت

$$G(x) = \exp\{-\exp[-\frac{x-a}{b}]\}$$

با پارامترهای $a < b < +\infty$ و $a < -\infty$ است. میانگین این توزیع، $a + b/5776$ و انحراف معیار آن $\sqrt{b/a}$ می‌باشد. در نتیجه، مدل لگ-لگ را می‌توان به صورت

از توزیع گامبل با $F(\pi(x)) = F(-x'\beta)$ و $a = 0$ و $b = 1$ در نظر گرفت.

۳.۳.۲ رویکرد متغیر پنهان

در این بخش، مدل‌هایی غیر خطی با استفاده از مفهوم متغیر پنهان معرفی می‌شوند. مشاهده خواهد شد که GLM‌ها را می‌توان با استفاده از این مدل‌ها نیز به دست آورد. در ابتدا متغیر پنهان تعریف می‌شود و سپس مدل‌ها در حالت کلی و جزئی بیان می‌شوند.

۱.۳.۳.۲ متغیر پنهان

متغیر Y^* را پنهان گوییم اگر مشاهده نشود و در فاصله‌ی $(-\infty, \infty)$ تغییر کند و Y ‌های مشاهده شده را به این صورت تولید کند که اگر Y^* مقادیر بزرگ اختیار کرد، $Y = 1$ و اگر Y^* مقادیر کوچک گرفت، $Y = 0$ مشاهده خواهد شد. کوچکی یا بزرگی Y^* توسط یک مقدار آستانه‌ای (معمولًاً 0) در نظر گرفته می‌شود.

سؤالی که مطرح می‌شود این است که آیا همیشه هر برآمد دودویی را می‌توان به صورت ظهور از یک متغیر پنهان در نظر گرفت؟ بعضی از محققان استدلال می‌کنند که در نظر گرفتن یک متغیر پنهان، معمولاً نامناسب است و بعضی دیگر براین عقیده‌اند که در همه‌ی حالت‌ها استفاده از مفهوم متغیر پنهان، معقول به نظر می‌رسد. صرف نظر از عقیده‌ی محققان در استفاده از متغیر پنهان، این مهم است که بدانیم به دست آوردن و کاربرد مدل‌های پاسخ دودویی، وابسته به پذیرفتن یا نپذیرفتن یک متغیر پنهان نیست.

در شکل ریاضی، متغیر Y^* مشاهدات دودویی Y را به صورت زیر تولید می‌کند:

$$y_i = \begin{cases} 1, & \text{اگر } \tau > Y^* \\ 0, & \text{اگر } \tau \leq Y^* \end{cases}$$

که در آن τ یک نقطه‌ی آستانه (قطع) می‌باشد. فرض می‌کنیم که $\tau = 0$. (کمی جلوتر نشان می‌دهیم که این فرض، یک فرض ضروری است و اکثر محققان، $\tau = 0$ را در نظر می‌گیرند).

به عنوان مثال، قصد خرید یک خودرو توسط سرپرست یک خانواده را در نظر بگیرید. اگر مجموع درآمد خانواده منهای مجموع هزینه‌ی خانواده (برای یک مدت زمان معین) یا

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

به عبارت دیگر، پس انداز خانواده، از یک مقدار معلوم بیشتر شود، سرپرست اقدام به خرید اتومبیل خواهد کرد و در غیر این صورت از خرید خودرو صرف نظر خواهد کرد. در این مثال، پس انداز خانواده را مشاهده نخواهیم کرد و فقط خریدن یا نخریدن خودرو توسط سرپرست خانواده را مشاهده می‌کنیم. بنا بر این، پس انداز خانواده یک متغیر پنهان است که متغیر مشاهده شده‌ی خریدن یا نخریدن خودرو را تولید می‌کند.

۲.۳.۳.۲ مدل‌های معرفی شده

فرض می‌شود متغیر پنهان Y^* که در زیربخش قبل معرفی شد، رابطه‌ی خطی با x های مشاهده شده به صورت زیر دارد:

$$Y^* = x'\beta + \varepsilon. \quad (9.2)$$

چون Y^* پیوسته است، مدل احتمالاتی خطی با آن مواجه است، اجتناب می‌کند. چون متغیر وابسته مشاهده نمی‌شود، مدل به وسیله‌ی کمترین توان‌های دوم معمولی براورد نمی‌شود. در عوض، از براورد ماکسیمم درست‌نمایی استفاده می‌کنیم که احتیاج به فرض‌هایی درباره‌ی توزیع خطاهای دارد. اغلب محققان، توزیع نرمال را (که مدل‌های پربویت را نتیجه می‌دهد) و توزیع لوژستیک را (که مدل‌های لوجیت را نتیجه می‌دهد) برای توزیع خطاهای انتخاب می‌کنند.

در مدل رگرسیون خطی فرض می‌کنیم که $E(\varepsilon|x) = 0$. چون y مشاهده نمی‌شود، واریانس آن قابل براورد نیست. در مدل پربویت فرض می‌شود $\text{var}(\varepsilon|x) = 1$ و در مدل لوجیت نیز $\text{var}(\varepsilon|x) = \pi^2/3 \approx 3/29$. این مقادیر خاص که برای واریانس فرض شده‌اند، دلخواه‌اند و مقادیری انتخاب می‌کنیم که ساده‌ترین فرمول برای توزیع ε را نتیجه دهد.

هنگامی که ε نرمال با $E(\varepsilon|x) = 0$ و $\text{var}(\varepsilon|x) = 1$ است، داریم:

$$\phi(\varepsilon) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\varepsilon^2}{2}\right\},$$

$$\Phi(\varepsilon) = \int_{-\infty}^{\varepsilon} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt, \quad (10.2)$$

که به ترتیب، تابع چگالی و تابع توزیع تجمعی ε هستند. در مدل لوجیت، خطاهای دارای توزیع لوژستیک با میانگین صفر و واریانس $\pi^2/3$ فرض

مدل‌های مختلف برای تحلیل داده‌های مقطعی دارای پاسخ دودویی

می‌شوند. با این میانگین و واریانس، تابع چگالی و تابع توزیع لوژستیک، به ترتیب به شکل زیرند:

$$\lambda(\varepsilon) = \frac{\exp(\varepsilon)}{[1 + \exp(\varepsilon)]^2},$$

$$\Lambda(\varepsilon) = \frac{\exp(\varepsilon)}{1 + \exp(\varepsilon)}. \quad (11.2)$$

تابع چگالی لوژستیک، پهن‌تر از تابع چگالی نرمال است؛ چون واریانس بیشتری دارد. اگر واریانس توزیع لوژستیک را مساوی با یک در نظر بگیریم، به آن توزیع لوژستیک استاندارد شده می‌گویند. آن‌گاه این توزیع و توزیع نرمال، تقریباً همسان می‌شوند، ولی تابع چگالی و تابع توزیع آن به ترتیب به صورت زیر در می‌آیند:

$$\lambda^s(\varepsilon) = \frac{\gamma \exp(\gamma\varepsilon)}{[1 + \exp(\gamma\varepsilon)]^2},$$

$$\Lambda^s(\varepsilon) = \frac{\exp(\gamma\varepsilon)}{1 + \exp(\gamma\varepsilon)}, \quad (12.2)$$

که در آن $\sqrt{\frac{\pi}{3}} = \gamma$ است. چون معادلات ساده‌تری برای تابع توزیع لوژستیک وجود دارد، عموماً از توزیع لوژستیک در فرمول (11.2) برای به دست آوردن مدل لوجیت استفاده می‌شود.

به وسیله‌ی مشخص کردن توزیع عها داریم:

$$\Pr(Y = 1|x) = \Pr(Y^* > 0|x) = \Pr(x'\beta + \varepsilon > 0|x) = \Pr(\varepsilon > -x'\beta|x)$$

چون توزیع‌های نرمال و لوژستیک متقارن‌اند، داریم:

$$\Pr(Y = 1|x) = \Pr(\varepsilon \leq x'\beta|x) = F(x'\beta),$$

که در آن، F تابع توزیع نرمال (Φ) برای مدل پربویت و تابع توزیع لوژستیک (Λ) برای مدل لوجیت است.

گیریم یک متغیر توضیحی داشته باشیم. در این صورت،

$$\Pr(Y = 1|x) = F(\alpha + \beta x).$$

افزایش یک واحد در x ، به این معنا است که $F^{-1}[\Pr(Y = 1|x)]$ به اندازه‌ی β واحد افزایش می‌یابد. اگر $x_1 > x_2$ باشد، $[F^{-1}[\Pr(Y = 1|x_1)] - F^{-1}[\Pr(Y = 1|x_2)]]$ به

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

اندازه‌ی $(x_1 - x_2)\beta$ تغییر می‌یابد. منحنی به دست آمده از ترسیم $(x, F(x'\beta))$, در صورت پیوسته بودن x , منحنی معروف S-شکل در مدل‌های دارای پاسخ دودویی است. در مشخص کردن مدل‌های پاسخ دودویی، سه فرض زیر را داریم:

(آ) $\tau = 0$, که در آن τ مقدار آستانه می‌باشد.

$$\text{ب) } E(\varepsilon|x) = 0$$

$$\text{پ) } \text{var}(\varepsilon|x) = \begin{cases} 1, & \text{در مدل پروبیت} \\ \frac{\pi^2}{3}, & \text{در مدل لوچیت} \end{cases}$$

این فرض‌ها اختیاری و دلخواه‌اند و نمی‌توانیم آن‌ها را آزمون کنیم، اما برای شناسایی مدل، لازم‌اند. در زیر، این مطلب مورد تأکید قرار می‌گیرد.

یکی از فرض‌هایی که برای شناسایی یک مدل لازم است و قبلًاً نیز ذکر شد، فرض $\tau = 0$ است. این فرض به دلیل زیرانتخاب می‌شود. فرض کنیم در معادله‌ی (۹.۲) فقط یک متغیر تبیینی داریم؛ یعنی معادله‌ی (۹.۲) به صورت $Y^* = \alpha + \beta x + \varepsilon$ خواهد بود. در

نتیجه:

$$\begin{aligned} \Pr(Y = 1|x) &= \Pr(Y^* > \tau|x) = \Pr(\alpha + \beta x + \varepsilon > \tau|x) \\ &= \Pr(\varepsilon > \tau - \alpha - \beta x|x) \\ &= \Pr(\varepsilon \leq \alpha - \tau + \beta x|x), \end{aligned}$$

در نتیجه:

$$\Pr(Y = 1|x) = F(\alpha - \tau + \beta x). \quad (۱۳.۲)$$

اگر بخواهیم پارامترهای مجھول معادله‌ی (۱۳.۲) را براورد کنیم، فقط $\tau - \alpha$ را می‌توانیم براورد کنیم و نه هر دو پارامتر α و τ را. به همین دلیل، یکی از آن‌ها را مقدار معلوم فرض می‌کنند. در اکثر موارد، $\tau = 0$ را اختیار می‌کنند تا α و β قابل براورد باشند.

چون متغیرهای پنهان مشاهده نمی‌شوند، میانگین و واریانس آن‌ها قابل براورد نیستند. در مدل‌های پاسخ دودویی، تا فرض‌هایی برای میانگین و واریانس Y^* وضع نشوند، مدل شناسایی نمی‌شود.

به دلیل زیر باید واریانس ϵ_y را مشخص کنیم. برای این منظور، رابطه‌ی بین واریانس متغیر وابسته و مشخصه‌ی β_y را که مدل رگرسیونی معمولی را بررسی می‌کنیم. مدل $Y^* = x'\beta_y + \epsilon_y$ را در نظر بگیرید. همان‌طور که فرض می‌شود، فقط ϵ_y مشاهده می‌شوند. یک متغیر وابسته‌ی جدید به صورت $W = \delta Y^*$ تعریف می‌کنیم، که δ یک ثابت غیر صفر است. در نتیجه برای $\delta \neq 0$ داریم:

$$\text{var}(W) = \text{var}(\delta Y^*) = \delta^2 \text{var}(Y^*).$$

اگر $\delta = 0$ باشد، داریم $\text{var}(W) = \frac{1}{\sqrt{\text{var}(Y^*)}}$

$$Y^* = x'\beta_y + \epsilon_y.$$

بنا بر این،

$$W = \delta(x'\beta_y + \epsilon_y) = x'(\beta_y\delta) + \delta\epsilon_y.$$

در نتیجه،

$$\beta_w = \delta\beta_y. \quad (14.2)$$

چون اندازه‌ی شیب، وابسته به پارامتر مقیاس متغیر وابسته است، اگر واریانس متغیر وابسته را ندانیم، ضرایب شیب مشخص نمی‌شوند.

این نتایج در مدل پاسخ دودویی و فهمیدن رابطه‌ی بین ضرایب مدل لوجیت در مقایسه با مدل پروبیت کاربرد دارد. برای توضیح بیشتر نیاز به تمییز دادن ساختار مدل برای لوجیت و پروبیت داریم.

گیریم $Y_L^* = x'\beta_L + \epsilon_L$ و $Y_P^* = x'\beta_P + \epsilon_P$ باشد، که در آن‌ها L شناسه‌ی مدل لوجیت و P شناسه‌ی مدل پروبیت است.

چون Y_P^* و Y_L^* متغیرهای پنهان هستند واریانس آن‌ها را نمی‌توان بر اساس مشاهدات تعیین کرد؛ در نتیجه β_L و β_P تشخیص داده نمی‌شوند. برای هر دو مدل، واریانس ϵ^* به وسیله‌ی فرض‌هایی برای واریانس ϵ تعیین می‌شود.

$$\text{var}(\epsilon_L|x) = (\frac{\pi}{\varphi}) \text{var}(\epsilon_P|x).$$

در نتیجه با توجه به (14.2) داریم $\epsilon_L \cong \frac{\pi}{\varphi} \epsilon_P$ ؛ یعنی:

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

$$\beta_L \simeq \sqrt{\text{var}(\varepsilon_L | x)} \beta_P$$

یا

$$\beta_L \simeq 1/\lambda \beta_P.$$

این تبدیل می‌تواند برای مقایسهٔ ضرایب از تحلیل مدل لوجیت در برابر مدل پروبیت و بر عکس، مورد استفاده قرار گیرد.

تقریب $\beta_L \simeq 1/\lambda \beta_P$ برپایهٔ واریانس‌های توزیع لوژستیک و نرمال به دست می‌آید. اممیا (۱۹۸۱) پیشنهاد می‌کند که توزیع‌های لوژستیک و نرمال را تا حدّ ممکن نزدیک کنیم و فقط به مساوی در نظر گرفتن واریانس‌ها اکتفا نکنیم. او می‌گوید هنگامی که $\epsilon_L \simeq 1/\lambda \epsilon_P$ باشد، این رابطه منجر به $\beta_L \simeq 1/\lambda \beta_P$ می‌شود و توزیع‌های آن‌ها خیلی شبیه به هم می‌شوند. گاهی $1/\lambda \epsilon_P$ را در نظر می‌گیرند.

چون β ‌ها بدون فرض‌هایی در بارهٔ میانگین و واریانس‌ها مشخص نمی‌شوند، β ‌ها در این حالت، اختیاری هستند. اگر فرض در بارهٔ $\text{var}(\varepsilon | x)$ را تغییر دهیم، β ‌ها نیز تغییر می‌کنند. بنا بر این β ‌ها به‌طور مستقیم نمی‌توانند تفسیر شوند، چون بیان‌کننده‌ی دو مورد می‌باشند:

(آ) ارتباط بین x ‌ها و y^*

(ب) فرض‌ها را مشخص می‌کنند.

مشخص کردن فرض‌ها بر $\Pr(Y = 1 | x)$ تأثیر ندارد؛ زیرا $\Pr(Y = 1 | x)$ تابعی قابل برآورده است. یک تابع قابل برآورده، تابعی از پارامترها است که نسبت به تغییر فرض‌ها پایا است (سیرل، ۱۹۷۱).

به عنوان مثال، مدل لوجیت را در نظر بگیرید:

$$\Pr(Y = 1 | x_i) = \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} = \frac{1}{1 + \exp(-x_i \beta)}.$$

سمت راست، تابع توزیع تجمعی برای توزیع لوژستیک با واریانس $\sigma^2 = \frac{\pi^2}{3}$ است. می‌توانیم ε را استاندارد کنیم تا دارای واریانس ۱ باشد. به وسیلهٔ مدل و ساختارهایی که به وسیلهٔ σ به دست می‌آیند، داریم:

$$\frac{y_i^*}{\sigma} = \frac{x_i \beta}{\sigma} + \frac{\varepsilon_i}{\sigma},$$

که ε_i دارای توزیع لوژستیک استاندارد شده با تابع توزیع تجمعی

$$\Lambda^s\left(\frac{\varepsilon_i}{\sigma}\right) = \frac{\exp\left(\frac{\pi \varepsilon_i}{\sqrt{3} \sigma}\right)}{1 + \exp\left(\frac{\pi \varepsilon_i}{\sqrt{3} \sigma}\right)} = \frac{\exp(\varepsilon_i)}{1 + \exp(\varepsilon_i)} = \Lambda(\varepsilon_i)$$

است. در نتیجه، تغییر فرض‌هایی در باره‌ی $\text{var}(\varepsilon|x)$ بر احتمال یک پیشامد، تأثیر نداشته است؛ یعنی این‌که مقدار خاص اختیاری برای $\text{var}(\varepsilon|x)$ روی β ‌ها اثر دارد، ولی بر کمیتی که مورد علاقه‌ی اساسی است (یعنی احتمال رخ دادن یک پیشامد)، اثر ندارد.

در نتیجه احتمال‌ها می‌توانند تفسیر شوند بدون آن که فرض‌هایی که برای ساختن یک مدل لازم است، بر روی آن‌ها تأثیر بگذارند. به این دلیل است که احتمال را تابعی قابل برآورد می‌گویند. در نتیجه تغییر در احتمال‌ها و بخت‌ها را، که نسبت احتمال موفقیت به احتمال شکست است، می‌توان تفسیر کرد.

۳.۳.۲.۲ صورت کلی مدل‌های ناخطي

در این بخش، یک صورت کلی از مدل‌های بالا ارائه می‌شود که توسط آراندال‌اورداز (۱۹۸۱) معرفی شده است. البته باید خوانندگان محترم توجه داشته باشند که این صورت، تنها صورت کلی ارائه شده نیست. در سال‌های اخیر، حالت‌های کلی دیگری هم ارائه شده‌اند که مدل‌های ذکر شده در بالا را تحت پوشش قرار می‌دهند؛ اما به دلیل پیچیده بودن، از ارائه‌ی آن‌ها صرف نظر می‌کنیم.

این شکل کلی GLM‌ها برای تحلیل پاسخ دودویی، توسط آراندال‌اورداز (۱۹۸۱) پیشنهاد شده است. دو خانواده از تبدیل‌های توانی برای احتمال‌ها معرفی می‌شوند که انحراف از مدل لوژستیک را به صورت متقارن و نامتقارن مدل‌بندی می‌کنند. در این قسمت، نمایشی کلی از GLM‌ها برای تحلیل داده‌های پاسخ دودویی مهیا می‌شود. تبدیل‌ها به دو قسمت متقارن و نامتقارن تقسیم می‌شوند (منظور از تبدیل‌ها، تبدیل‌های توانی روی احتمال‌ها است که به صورت یک تابع خطی به متغیرهای تبیینی وابسته است).

آ) تبدیل‌های متقارن

همان‌طور که ذکر شد، تبدیل لوجیت، یکی از تبدیل‌های متقارن است. تابع لوجیت به صورت زیر تعریف شده است:

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

آرندالورداز (۱۹۸۱) خانواده‌ی زیر از تبدیل‌ها را در نظر گرفته‌اند:

$$T_\lambda(\pi) = \frac{2}{\lambda} \cdot \frac{\pi^\lambda - (1-\pi)^\lambda}{\pi^\lambda + (1-\pi)^\lambda}, \quad (15.2)$$

که در آن، λ پارامتر تبدیل نامیده می‌شود.

دو خاصیت ساده از (۱۵.۲) عبارت‌اند از

$$T_\lambda(\theta) = -T_\lambda(1-\theta), \quad T_\lambda(\theta) = T_{-\lambda}(\theta);$$

یعنی این‌که T_λ ، موفقیت و شکست را به صورت متقارن مورد بررسی قرار می‌دهد. بنا بر این، خانواده‌ی تبدیل‌های رابطه‌ی (۱۵.۲) را تبدیل‌های متقارن می‌نامند.

معادله‌ی (۱۵.۲) هنگامی که $0 \rightarrow \lambda$ ، به لوجیت و هنگامی که $1 = \lambda$ ، به مدل احتمالاتی خطی (ضرب در یک مقدار ثابت) تبدیل می‌شود.

در این حالت، فرض می‌کنند $x'\beta = T_\lambda(\pi)$; یعنی این‌که $T_\lambda(\pi)$ بر حسب عبارت $x'\beta$ با متغیرهای تبیینی مرتبط می‌شود. حال اگر درست‌نمایی (لگاریتم درست‌نمایی) به صورت تابعی از λ در نظر گرفته و مدل به وسیله‌ی روش ماکسیمم درست‌نمایی به‌ازای مقادیری از λ برآش داده شود، نه تنها براوردگر ماکسیمم درست‌نمایی λ (یعنی $\hat{\lambda}$) به دست می‌آید، بلکه مقدارهایی از λ را می‌توان به دست آورد که یک برآش قابل قبول می‌دهند.

ب) تبدیل‌های نامتقارن

همان‌طور که قبلاً نیز بیان شد، حالت‌هایی وجود دارند که در آن‌ها موفقیت و شکست باید به صورت نامتقارن بحث شوند. در این حالت، خانواده‌ی تبدیل‌های زیر در نظر گرفته می‌شود:

$$W_\lambda(\pi) = \frac{\{(1-\pi)^{-\lambda} - 1\}}{\lambda}.$$

پس:

$$T_\lambda(\pi) = \log(W_\lambda(\pi)), \quad (16.2)$$

که (۱۶.۲) همانند قسمت قبل به صورت $x'\beta = T_\lambda(\pi)$ به متغیرهای تبیینی وابسته است. رابطه‌ی (۱۶.۲) به‌ازای $1 = \lambda$ به مدل لوژستیک وقتی $0 \rightarrow \lambda$ به مدل لگ-لگ

مکمل تبدیل می‌شود. بنا بر این، این دو مدل را فقط به وسیله‌ی یک پارامتر λ می‌توان مقایسه کرد.

با توجه به این دو تبدیل، مشاهده می‌شود که اگرتابع ربط، $(\pi)T_\lambda$ در نظر گرفته شود یک GLM به دست می‌آید. در نتیجه با براورد کردن λ به روش ماکسیمم درست‌نمایی توسط خود داده‌ها، می‌توان تابع ربط مناسبی در GLM برای داده‌های دودویی به دست آورد (خود داده‌ها مقدار λ را مشخص می‌کنند).

۴.۲ روش‌های براورد

۱.۴.۲ روش شبه‌درست‌نمایی

شبه‌درست‌نمایی، اولین بار توسط ودربرن (۱۹۷۴) و سپس به طور وسیع‌تری به وسیله‌ی مکولا (۱۹۸۳) مورد تحقیق قرار گرفته است.

شبه‌درست‌نمایی، روشی برای براورد کردن پارامترهای رگرسیونی است که به فرض‌های کمکی درباره‌ی توزیع متغیر وابسته احتیاج دارند و بنا بر این با برآمدهای گوناگون می‌تواند مورد استفاده قرار گیرد. در تحلیل درست‌نمایی، شکل توزیع باید مشخص شود؛ ولی در شبه‌درست‌نمایی، فقط روابط بین میانگین برآمد و متغیرهای کمکی و نیز رابطه‌ی بین میانگین و واریانس باید مشخص شود.

برای داده‌های دودویی در GLM‌ها با تابع ربط لوجیت، ضرایب رگرسیونی می‌توانند به وسیله‌ی حل کردن معادلات براوردگر زیر براورد شوند:

$$S(\beta) = \sum_{i=1}^n \left(\frac{\partial \pi_i}{\partial \beta} \right)' v_i^{-1} \{y_i - \pi_i\} = 0, \quad (17.2)$$

که در آن‌ها $y_i = \text{var}(Y_i)$ و $S(\beta)$ مشتق لگاریتم تابع درست‌نمایی است. برای درک بیشتر این مطلب، همان‌طور که در زیربخش ۲.۴.۲ خواهیم دید، برای تابع ربط لوجیت داریم:

$$\ell_i = y_i \log \left(\frac{\pi_i(x)}{1 - \pi_i(x)} \right) + \log(1 - \pi_i(x)).$$

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

در نتیجه اگر $\theta_i = \text{logit}(\pi_i(x))$ در نظر گرفته شود، خواهیم داشت:

$$\log(1 - \pi_i(x)) = -\log(1 + \exp(\theta_i)).$$

همچنین اگر برای ساده‌نویسی، $\pi_i(x)$ را با π_i نشان دهیم، خواهیم داشت:

$$\frac{\partial \ell_i}{\partial \beta} = \left(\frac{\partial \pi_i}{\partial \beta} \right)' \frac{\partial \theta_i}{\partial \pi_i} \frac{\partial \ell_i}{\partial \theta_i},$$

که بعد از انجام دادن مشتق‌گیری‌ها خواهیم داشت:

$$\frac{\partial \ell_i}{\partial \beta} = \left(\frac{\partial \pi_i}{\partial \beta} \right)' v_i^{-1} \{y_i - \pi_i\}.$$

در نتیجه معادلات (۱۷.۲) به دست می‌آیند. بنا براین معادلات (۱۷.۲) در حالت استفاده از ربط لوجیت، معادلات درست‌نمایی هستند.

براورد $(\hat{\beta})$ ، که از حل معادلات (۱۷.۲) به دست می‌آید، دارای توزیع مجانبی گاوی با میانگین β و واریانس زیر خواهد بود (دیگل و دیگران، ۲۰۰۲):

$$V = \left\{ \sum_{i=1}^n \left(\frac{\partial \pi_i}{\partial \beta} \right)' v_i^{-1} \left(\frac{\partial \pi_i}{\partial \beta} \right) \right\}^{-1} \times \left(\sum_{i=1}^n \left(\frac{\partial \pi_i}{\partial \beta} \right)' v_i^{-1} \{Y_i - \pi_i(\beta)\} v_i^{-1} \left(\frac{\partial \pi_i}{\partial \beta} \right) \right)^{-1} \\ \left\{ \sum_{i=1}^n \left(\frac{\partial \pi_i}{\partial \beta} \right)' v_i^{-1} \left(\frac{\partial \pi_i}{\partial \beta} \right) \right\}^{-1}. \quad (۱۸.۲)$$

خاصیت مهمی از GLM‌ها، که در رابطه‌ی (۱۷.۲) مشاهده می‌شود، این است که تابع امتیاز، $S(\beta)$ ، فقط به میانگین و واریانس Y_i ‌ها وابسته است. بنا براین و در برن (۱۹۷۴) بیان کرد که از معادلات براورده (۱۷.۲) می‌توان به منظور براورد کردن ضرایب رگرسیونی برای انتخاب هر نوع تابع ربط و تابع واریانس استفاده کرد (صرف نظر از این که آیا لذا عضو خانواده‌ی نمایی تک‌پارامتری هستند یا نه). این منجر به رهیافتی برای مدل‌بندی آماری می‌شود که فقط نیاز به فرض‌هایی درباره‌ی تابع‌های ربط و واریانس دارد، بدون آن که فرضی درباره‌ی توزیع احتمالی شود که مشاهدات از آن حاصل شده‌اند، داشته باشند.

شبه‌درست‌نمایی، رهیافت خوبی به نظر می‌رسد؛ چون غالباً مکانیسم احتمالاتی ای را که داده‌ها از آن تولید شده‌اند، نمی‌توان به درستی مشخص کرد.

پس اگر فرض کنیم $\pi_i = x_i' \beta$ و $\text{var}(Y_i) = v(\pi_i)$ ، که در آن v تابعی معلوم از π است، براورد ضرایب رگرسیونی (β) جواب معادلات (۱۷.۲) است. و در برن (۱۹۷۴) نشان داد هنگامی که توزیع مشاهدات، خانواده‌ی نمایی تک‌پارامتری نرمال است، شبیدرستنمایی، همان لگاریتم درستنمایی خواهد بود.

مکولا (۱۹۸۳) نشان داد جواب معادله‌ی شبیدرستنمایی $(\hat{\beta})$ با تابع ربط لوجیت که از حل معادله‌ی (۱۷.۲) به دست می‌آید، دارای توزیع مجانبی گاووسی با میانگین β و واریانس داده شده در معادله‌ی (۱۸.۲) خواهد بود.

۴.۴.۲ معادلات درستنمایی و براوردهای ماکسیمم درستنمایی

معادلات درستنمایی را به وسیله‌ی تعریف کردن p_i برای n مشاهده، به صورت زیر به

دست می‌آوریم:

$$p_i = \begin{cases} \Pr(Y_i = 1|x_i), & \text{اگر } Y_i = 1 \text{ مشاهده شود} \\ 1 - \Pr(Y_i = 1|x_i), & \text{اگر } Y_i = 0 \text{ مشاهده شود} \end{cases}$$

$$\pi_i(x) = \Pr(Y_i = 1|x_i) = F(x_i' \beta)$$

پس با فرض مستقل بودن مشاهدات، معادلات درستنمایی عبارت‌اند از

$$\begin{aligned} L(\beta; y, x) &= \prod_{i=1}^n p_i \\ &= \prod_{\{i: Y_i = 1\}} \Pr(Y_i = 1|x_i) \prod_{\{i: Y_i = 0\}} [1 - \Pr(Y_i = 1|x_i)] \\ &= \prod_{\{i: Y_i = 1\}} F(x_i' \beta) \prod_{\{i: Y_i = 0\}} [1 - F(x_i' \beta)]. \end{aligned}$$

در نتیجه:

$$\ell = \log[L(\beta; y, x)] = \sum_{\{i: Y_i = 1\}} \log[F(x_i' \beta)] + \sum_{\{i: Y_i = 0\}} \log[1 - F(x_i' \beta)].$$

به عنوان مثال، اگر مدل لوجیت را در نظر بگیریم، یعنی

$$\log \left(\frac{\pi_i(x)}{1 - \pi_i(x)} \right) = x_i' \beta = \sum_j x_{ij} \beta_j,$$

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

تابع درست‌نمایی به صورت زیر خواهد بود:

$$\begin{aligned} L(\beta; y, x) &= \prod_{i=1}^n [\pi_i(x)^{y_i} (1 - \pi_i(x))^{1-y_i}] \\ &= \prod_{i=1}^n \left[(1 - \pi_i(x)) \left(\frac{\pi_i(x)}{1 - \pi_i(x)} \right)^{y_i} \right] \\ &= \prod_{i=1}^n \exp \left\{ \log(1 - \pi_i(x)) + y_i \log \left(\frac{\pi_i(x)}{1 - \pi_i(x)} \right) \right\}. \end{aligned}$$

می‌دانیم که

$$\log \left(\frac{\pi_i(x)}{1 - \pi_i(x)} \right) = \sum_j x_{ij} \beta_j.$$

در نتیجه:

$$(1 - \pi_i(x)) = [1 + \exp(\sum_j x_{ij} \beta_j)]^{-1}.$$

پس معادله‌ی درست‌نمایی به صورت زیر در می‌آید:

$$L(\beta; y, x) = \exp \left\{ - \sum_{i=1}^n \left[\log(1 + \exp(\sum_j x_{ij} \beta_j)) - y_i \sum_j x_{ij} \beta_j \right] \right\}.$$

اگر از معادله‌ی بالا لگاریتم بگیریم، داریم:

$$\ell(\beta) = \log\{L(\beta; y, x)\} = \sum_j (\sum_{i=1}^n x_{ij} y_i) \beta_j - \sum_{i=1}^n \log\{1 + \exp(\sum_j \beta_j x_{ij})\}.$$

اممیا (۱۹۸۱) ثابت کرد که تحت شرایطی که در بیش‌تر موارد برقرارند، تابع درست‌نمایی مقعر می‌شود و این دلیل بریکتاپی براوردگر ماکسیمم درست‌نمایی است. این براوردگرها سازگار، به‌طور مجانبی نرمال، و به‌طور مجانبی کارا می‌باشند.

برای مدل‌های ناخطي، براوردگرهای ماکسیمم درست‌نمایی از به دست آوردن مشتق L (یا ℓ) نسبت به β و برابر با صفر قرار دادن آن به‌طور جبری، به دست نمی‌آیند و باید از روش‌های عددی برای به دست آوردن براوردها استفاده شود. امروزه در بیش‌تر نرم‌افزارهای آماری، رگرسیون با پاسخ دودویی برای توابع ربط گوناگون موجود است (چگونگی استفاده از نرم‌افزار R را در زیربخش ۶.۲ ببینید).

پس از برازش مدل، مسئله‌ی بعدی، بررسی مانده‌ها است. مانده‌های خام، اختلاف بین مقادیر برازandه شده و مشاهده شده را نشان می‌دهند. در داده‌های دوچمله‌ای، این اختلاف

۵.۲ مثال‌های کاربردی

۶۹

برابر است با $y_i - n_i \hat{\pi}_i$. معمولاً مانده‌های خام در مقابل متغیرهای تبیینی رسم می‌شوند. برای یک مدل مناسب، نباید روندی در چنین نمودارهایی دیده شود. به هر حال، چون مانده‌های خام، استاندارد شده نیستند، از آن‌ها استفاده‌ی اندکی برای تشخیص نقاط دورافتاده می‌شود.

مانده‌های پیرسون و مانده‌های کیبس به صورت زیر تعریف می‌شوند:

$$r_{p_i} = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}},$$

$$r_{D_i} = \text{sign}(y_i - n_i \hat{\pi}_i) \left[2y_i \log \left(\frac{y_i}{n_i \hat{\pi}_i} \right) + 2(n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - n_i \hat{\pi}_i} \right) \right]^{\frac{1}{2}}.$$

اگر π_i ‌ها نزدیک به ۰ یا ۱ نباشند، هر دو این مانده‌ها دارای واریانس تقریبی ۱ هستند. مانده‌های استاندارد شده نیز عبارت‌اند از

$$r_{p_i}^* = \frac{r_{p_i}}{\sqrt{1 - H_{ii}}},$$

$$r_{D_i}^* = \frac{r_{D_i}}{\sqrt{1 - H_{ii}}},$$

که در آن‌ها H_{ii} درایه‌های روی قطر ماتریس کلاهدار H هستند. ماتریس کلاهدار H به صورت $W^{\frac{1}{2}} X (X' W X)^{-1} X' W^{\frac{1}{2}}$ تعریف می‌شود که در آن W ماتریس قطری با عناصر $\frac{(\frac{\partial \pi_i}{\partial \eta_i})^2}{\text{var}(Y_i)}$ است و در حالتی که پاسخ دودویی باشد $\sum_j \beta_j x_{ij} = \log(\frac{\pi_i}{1 - \pi_i})$ است. $r_{p_i}^* = \eta_i$ و $r_{D_i}^*$ نیز برای نمونه‌های بزرگ، واریانس مجانبی ۱ دارند. مقادیر بزرگ قدر مطلق این مانده‌ها (بزرگ‌تر از ۳)، نقاط دورافتاده را مشخص می‌کنند.

۵.۲ مثال‌های کاربردی

۱.۵.۲ اثر دوز دیسولفید کربن

همان‌طور که در فصل اول به آن اشاره شد، مثالی از پاسخ‌های مقطعی دودویی، بررسی اثر دوزهای مختلف دیسولفید کربن بر روی سوسک‌ها بود، که متغیر پاسخ در این مثال، دو حالت زنده ماندن و مردن سوسک‌ها بود. در جدول ۲.۲، تعداد سوسک‌های کشته شده بعد از ۵ ساعت قرار گرفتن در معرض گاز دیسولفید کربن در غلظت‌های مختلف، ارائه شده

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

است. اکنون با برآزندن مدل‌های لوجیت، پروبیت و لگ-لگ مکمل که در این بخش به عنوان مدل‌های مناسب برای پاسخ‌های دودویی معرفی شدند، مقادیر برآزانده شده را به دست می‌آوریم و با هم مقایسه می‌کنیم.

برازش ماکسیمم درست‌نمایی مدل پروبیت به صورت زیر است:

$$\Phi^{-1}[\hat{\pi}(x)] = -\frac{34}{96} + \frac{19}{74}x.$$

برای این برازش، در نقطه‌ی $\frac{34}{96} = x$ داریم $0/5 = \hat{\pi}(x)$. در مقدار دوز x_i با n_i سوسک، $(n_i \hat{\pi}(x_i))$ تعداد سوسک‌های مرده‌ی برآزانده شده به ازای $i = 1, \dots, 8$ است. جدول ۲.۲ مقادیر برآزانده شده برای این مدل را نشان می‌دهد. این جدول، همچنین مقادیر برآزانده شده برای مدل لوجیت را نیز نشان می‌دهد. این مدل‌ها به‌طور مشابه و ضعیف برآورد شده‌اند. آماره‌ی نیکویی برازش خی دو که در فصل اول معرفی شد، در این مثال، مقدار $11/1$ برای مدل لوجیت و مقدار $1/10$ برای مدل پروبیت با 6 درجه آزادی است.

جدول ۲.۲: برآورد پارامترهای مدل‌های مختلف برای داده‌های اثر دوز بر دیسولفید کربن

لوجیت	مقادیر برآزانده شده	لگ-لگ مکمل	تعداد سوسک‌های کشته شده	تعداد سوسک‌ها	تعداد سوسک‌ها	لگاریتم دوز
	پروبیت					
$2/5$	$2/4$	$5/7$	6	59	59	$1/691$
$9/8$	$10/7$	$11/3$	13	60	60	$1/724$
$22/4$	$22/4$	$20/9$	18	62	62	$1/755$
$32/9$	$32/8$	$30/3$	28	56	56	$1/784$
$50/0$	$49/6$	$42/7$	52	63	63	$1/811$
$52/3$	$52/4$	$54/2$	52	59	59	$1/837$
$59/2$	$59/7$	$61/1$	61	62	62	$1/861$
$58/8$	$59/2$	$59/9$	60	60	60	$1/884$

مدل لگ-لگ مکمل نیز به این داده‌ها برازش داده شده است، که برآورد ماکسیمم درست‌نمایی پارامترهای آن، $\hat{\alpha} = 22/01 = 0.22$ و $\hat{\beta} = -39/52 = -0.76$ است. به ازای دوز $x = 1/7$ ، احتمال برآزانده شده‌ی زنده ماندن برابر است با

$$1 - \hat{\pi}(x) = \exp\{-\exp[-39/52 + 22/01(1/7)]\} = 0.885,$$

در حالی که این کمیت به ازای $x = 1/8$ برابر است با 0.332 و در نقطه‌ی $x = 1/9$ برابر با $0.105 \times 5 = 0.5$ است. نتایج جدول ۲.۲ نشان می‌دهد که مقادیر برآزانده شده توسط

مدل لگ-لگ مکمل، به مقادیر مشاهده شده نزدیک‌تر از دو مدل معرفی شده‌ی دیگر است. آماره‌ی خی دو برای این مدل، برابر با $3/5$ است که تأییدکننده‌ی این نکته است.

۲.۵.۲ بررسی اثر پرتودهی گاز دی‌اکسید نیتروژن و زمان پرتودهی بر احتمال مرگ موش‌ها

همان‌طور که در فصل اول به آن اشاره شد، از دیگر مثال‌های مربوط به مطالعات مقطعی با پاسخ دودویی، مثال مربوط به موش‌ها است که هدف از آن، بررسی اثر پرتودهی گاز دی‌اکسید نیتروژن و زمان پرتودهی بر احتمال مرگ موش‌ها بود. با توجه به این‌که پیشنهاد شده است که از لگاریتم هر یک از متغیرهای کمکی استفاده شود، با انجام دادن این تبدیل، از متغیرهای کمکی جدید در مدل‌بندی استفاده می‌شود. سپس هر متغیر کمکی را استاندارد می‌کنیم و تأثیر متغیرهای کمکی استاندارد شده را بر نسبت بخت می‌سنجدیم. در این مثال، در ابتدا با استفاده از مدل لوچیت به بررسی اثرهای اصلی متغیرهای کمکی می‌پردازیم و سپس مدلی را معرفی می‌کنیم که علاوه بر اثرهای اصلی، اثرهای متقابل متغیرهای کمکی را نیز در نظر می‌گیرد. مدل $\log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} = 0/026 + 1/012x_{i1} + 1/0131x_{i2}$ مدل رگرسیون لوچیت برآورد شده به داده‌های مربوط به این مسئله است، که فقط اثرهای اصلی را در نظر می‌گیرد. در این مدل، x_1 درجه‌ی پرتودهی و x_2 زمان پرتودهی است. مدل $\log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} = 0/185 + 1/038x_{i1} + 1/237x_{i2} + 0/229x_{i1}x_{i2}$ مدل رگرسیون لوچیت برآورد شده با در نظر گرفتن اثرهای متقابل، علاوه بر اثرهای اصلی است. نتایج این مدل، نشان‌دهنده‌ی تأثیرگذاری متقابل درجه‌ی پرتودهی و زمان پرتودهی است و همچنین همبستگی قوی با احتمال مرگ را نشان می‌دهد.

به منظور بررسی مناسبت و نیکویی برآش مدل، آماره‌ی نسبت درست‌نمایی را معرفی می‌کنیم که با کیبیش مانده‌ها نمایش داده می‌شود. این آماره برای پاسخ‌های دودویی با استفاده از رابطه‌ی زیر به دست می‌آید:

$$\begin{aligned} LRS &= 2 \sum_{i=1}^m y_i \log y_i - 2 \sum_{i=1}^m y_i \log \hat{y}_i + 2 \sum_{i=1}^m (n_i - y_i) \log(n_i - y_i) \\ &\quad - 2 \sum_{i=1}^m (n_i - y_i) \log(n_i - \hat{y}_i), \end{aligned}$$

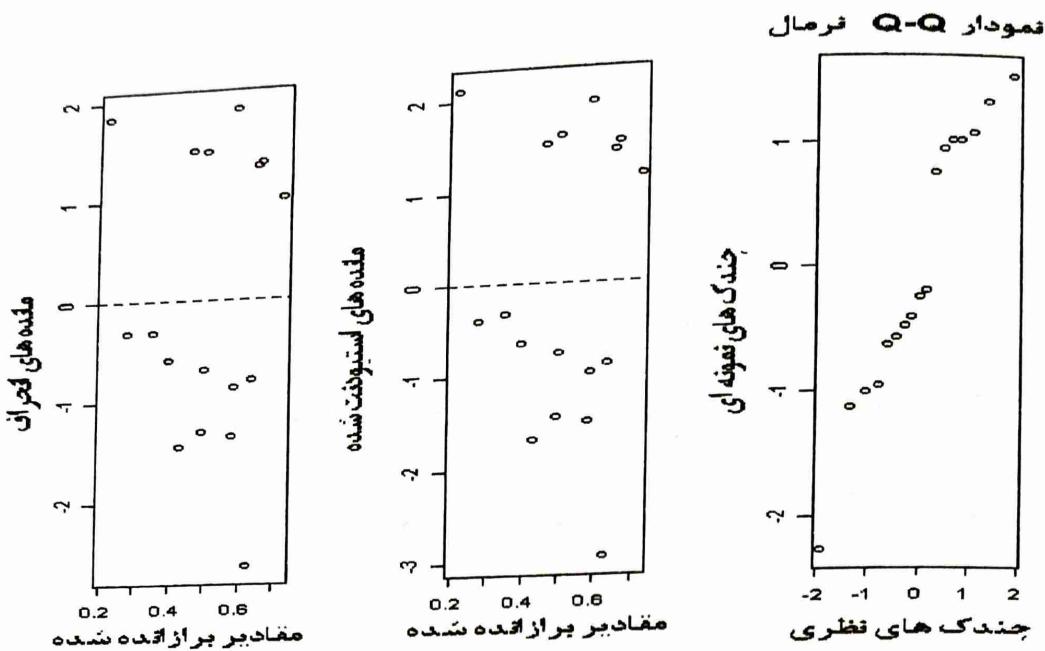
فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

که در آن $n_i \hat{\pi}_i = \hat{y}_i$ و $\hat{\pi}_i$ مقدار برازانده شده‌ی π_i است. مقدار این آماره با سطح بالای توزیع خی دو با $p - m$ درجه‌ی آزادی در این نقطه مقایسه می‌شود، که در آن m تعداد گروه‌ها و p تعداد پارامترهای مدل مورد استفاده است.

جدول ۳.۲: مقادیر مانده‌ها، احتمال و لگاریتم نسبت بخت‌های پیش‌بینی شده در داده‌های مربوط به موش‌های در معرض پرتودهی NO_2 با استفاده از برازش مدل با اثرهای اصلی و اثرهای متقابل

شماره‌ی گروه	تعداد موش‌های مرده (y_i)	تعداد کل موش‌ها (n_i)	مانده‌ها	احتمال مرگ	لگاریتم نسبت بخت
۱	۴۴	۱۲۰	-۱/۴۴	۰/۴۳۱	-۰/۲۷۵
۲	۳۷	۸۰	-۰/۶۷۹	۰/۵۰۰	۰/۰۰۲
۳	۴۳	۸۰	-۰/۸۷۹	۰/۵۸۵	۰/۳۴۶
۴	۳۵	۷۰	-۰/۷۹۹	۰/۶۲۳	۰/۵۴۷
۵	۲۹	۱۰۰	۱/۸۴۴	۰/۲۱۱	-۱/۲۱۴
۶	۵۳	۲۰۰	-۰/۳۰۸	۰/۲۷۴	-۰/۹۷۰
۷	۱۳	۴۰	-۰/۳۱۰	۰/۲۴۸	-۰/۶۲۷
۸	۷۵	۲۰۰	-۰/۵۸۵	۰/۳۹۵	-۰/۴۲۵
۹	۲۲	۴۰	۱/۴۹۳	۰/۴۰۷	-۰/۱۷۲
۱۰	۱۵۲	۲۸۰	۱/۴۷۷	۰/۴۹۸	-۰/۰۰۵
۱۱	۵۵	۸۰	۱/۹۰۹	۰/۵۸۳	۰/۳۳۹
۱۲	۹۸	۱۴۰	۱/۳۲۵	۰/۶۴۷	۰/۶۰۶
۱۳	۱۲۱	۱۶۰	۱/۰۰۱	۰/۷۲۱	۰/۹۵۰
۱۴	۵۲	۱۲۰	-۱/۲۹۷	۰/۴۹۲	-۰/۰۳۰
۱۵	۶۲	۱۲۰	-۱/۳۴۹	۰/۵۷۷	۰/۳۱۳
۱۶	۶۱	۱۲۰	-۲/۶۲۲	۰/۶۲۶	۰/۵۱۵
۱۷	۸۶	۱۲۰	۱/۳۵۸	۰/۶۵۸	۰/۶۵۷

با توجه به این که نتایج برازش مدل، صرفاً با در نظر گرفتن اثرهای اصلی، مقدار آماره‌ی نسبت درست‌نمایی را برابر با $۳۰/۹۳$ با ۱۴ درجه‌ی آزادی نشان می‌دهد، نیکویی برازش این مدل رد می‌شود؛ یعنی، مدلی که فقط از اثرهای اصلی در برازش مدل استفاده کرده است، مدل مناسبی به نظر نمی‌رسد. با اضافه کردن عبارت اثرهای متقابل، آماره‌ی کیبیش مانده‌ها به $۱۶/۱۱$ با ۱۳ درجه‌ی آزادی کاهش یافته است. مقدار مانده‌ها، $(y_i - \hat{y}_i) = (y_i - n\hat{\pi}_i)$ ، احتمال و لگاریتم نسبت بخت‌های پیش‌بینی شده نیز قابل محاسبه‌اند، که در این مثال، این مقادیر در جدول ۳.۲ برای ۱۷ گروه موش مورد بررسی، ارائه شده‌اند. همچنین به منظور کنترل فرض‌های مدل و بررسی انحراف از مدل، می‌توان از نمودارهای مانده‌ها استفاده کرد. نمودارهای دیگری از جمله نمودار Q-Q نرمال مانده‌ها برای مشخص کردن نقاط دورافتاده و رفتار داده‌ها وجود دارد که شکل ۱.۲ برای داده‌های موش‌ها، به نمودار مانده‌های استیوونت شده و مانده‌های انحراف نسبت به مقادیر برازانده شده و همچنین نمودار Q-Q نرمال اشاره دارد.



شکل ۱.۲: نمودار ماندها و نمودار $Q-Q$ نرمال بر اساس برازش مدل با اثرهای اصلی در داده‌های موش‌ها

۶.۲ دستورهای R برای برازش مدل در داده‌های مقطعی دارای پاسخ دودویی

در این بخش با استفاده از دستورهای مورد استفاده در نرم‌افزار R به تحلیل داده‌های مربوط به موش‌ها (مثال در زیربخش ۲.۵.۲) می‌پردازیم (داده‌ها در پیوست ۱.۲ داده شده است). با استفاده از دستور `read.table` در نرم‌افزار R می‌توان داده‌ها را که به صورت فایل txt ذخیره شده‌اند، فراخوانی کرد. همان‌طور که در مثال موش‌ها، در ابتدا با انجام پیشنهاد استفاده از لگاریتم هر یک از دو متغیر کمکی در مثال موش‌ها، در ابتدا با انجام دادن این تبدیل، متغیرهای کمکی جدید را به دست می‌آوریم. سپس هر متغیر کمکی را استاندارد کرده، تأثیر متغیرهای کمکی استاندارد شده را بر نسبت بخت‌ها می‌سنジم. دستورهای زیر برای ورود داده‌ها و انجام دادن تبدیل‌های فوق است.

```
data=read.table("mice.txt",header=T)
no2=data$no2
time=data$time
```

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

```

y=data$y
n=data$n
logno2=log(no2)
logtime=log(time)
logno2=(logno2-mean(logno2))/sqrt(var(logno2))
logtime=(logtime-mean(logtime))/sqrt(var(logtime))

```

قبل از مدل‌بندی، لازم است دو بردار y و $n - y$ به صورت یک ماتریس با دو ستون با استفاده از دستور زیر تولید شوند:

```
combdat=cbind(y, nminusy=n-y)
```

سپس با استفاده از دستور `glm` و صرفاً در نظر گرفتن اثرهای اصلی متغیرهای کمکی در مدل، به برازش پارامترهای مدل می‌پردازیم.

```

model1=glm(combdat ~ logno2+logtime, family=binomial(link=logit),
data=data)
print(summary(model1))

```

مدل ۱ که برای خانواده‌ی دوجمله‌ای تعریف شده است، از تابع ربط لوجیت `link=logit` استفاده کرده که به عنوان پیش‌فرض دستور `glm` است. توابع ربط دیگری چون `cloglog` و `probit` و `identity` و `inverse` و `square` و `inverse` نیز وجود دارند که می‌توان از آن‌ها نیز استفاده کرد. نتایج مدل ۱ که در شکل ۲.۲ آمده، نشان‌دهنده‌ی معنادار بودن دو متغیر کمکی در مدل است. به منظور دست‌یابی به مانده‌ها برای داده‌های دودویی، در نرم‌افزار R با استفاده از دستور `(.) residuals` امکان انتخاب سه نوع از مانده‌ها، مانده‌های انحراف (`type="deviance"`)، مانده‌های پی‌بررسونی (`type="pearson"`)، و مانده‌های عملی (`type="working"`) فراهم می‌شود. به منظور تحلیل‌های تشخیصی، مانده‌های انحراف (پیش‌فرض نرم‌افزار R)، ترجیح داده می‌شود.

```

Call:
glm(formula = combdat ~ logno2 + logtime, family = binomial
(link = logit), data = data)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-2.6223 -0.8699 -0.3100  1.3580  1.9098 

Coefficients:
            Estimate Std. Error   z value Pr(>|z|)    
(Intercept) 0.02686   0.04623   0.581   0.561    
logno2       1.01213   0.09025  11.215 <2e-16 ***  
logtime      1.13119   0.09302  12.161 <2e-16 ***  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 200.090 on 16 degrees of freedom
Residual deviance: 30.926 on 14 degrees of freedom
AIC: 122.47

Number of Fisher Scoring iterations: 3

```

شکل ۲.۲: نتایج برازش مدل ۱ بدون در نظر گرفتن اثرهای متقابل

مقدار ماندها، احتمال و لگاریتم نسبت بخت‌های پیش‌بینی شده، به ترتیب با استفاده از دستورهای `predict.glm` و `fitted.values` و `residuals` محاسبه می‌شوند. شایان ذکر است که اگر در دستور `predict.glm` عبارت `type="response"` را اضافه کنیم، به همان نتایج مربوط به دستور `fitted.values` دست می‌یابیم که براورد مقدار احتمال مرگ است؛ در حالی که اگر از این دستور اضافی استفاده نکنیم، براورد مقدار لگاریتم نسبت بخت‌ها را ارائه می‌کند. همان‌طور که قبلاً اشاره شد، به منظور کنترل فرض‌های مدل و بررسی انحراف از مدل، می‌توان از نمودارهای مانده‌ها استفاده کرد. در زیر به دستورهای به کار رفته برای دست‌یابی به این نمودارها اشاره شده است. نمودارهای دیگری از جمله نمودار Q-Q مانده‌ها (برای مشخص کردن نقاط دورافتاده و رفتار داده‌ها) که در کتابخانه‌ی MASS موجود است، با استفاده از دستورهای زیر، قابل حصول است:

```

residuals(model1)
rstudent(model1)
fitted.values(model1)

```

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

```

predict.glm(model1)
anova(model1)
par(mfrow=c(1,2))
plot(fitted.values(model1), residuals(model1))
abline(h=0,lty=2)
plot(fitted.values(model1), rstudent(model1))
abline(h=0,lty=2)
library(MASS)
qqnorm(studres(model1))

```

با توجه به این‌که نتایج مدل ۱، برازش خوبی را برای این داده‌ها نشان نداده است، دستورهای زیر، برازش مدلی را ارائه می‌کنند که در آن علاوه بر اثرهای اصلی، اثرهای متقابل متغیرهای کمکی نیز در نظر گرفته شده است:

```

model2=glm(combdat~ logno2*logtime, family=binomial
(link=logit), data=data)
print(summary(model2))

```

نتایج شکل ۳.۲ نشان‌دهنده‌ی تأثیرگذاری متقابل درجه‌ی پرتودهی و زمان پرتودهی است و همچنین همبستگی قوی با احتمال مرگ را نشان می‌دهد.

همان‌طور که قبلاً اشاره شد، یکی از روش‌های محاسباتی برای تحلیل داده‌ها روش شبه‌درست‌نمایی است. اگر مقادیر y و $y - n$ در جدول ۳.۲ را که به ترتیب، تعداد پاسخ‌های ۱ و صفر هستند، به صورت برداری از مقادیر ۱ و ۰ با نام جدید y_{new} نمایش دهیم و همچنین متغیرهای کمکی مربوط به هر یک از پاسخ‌های صفر و یک را با نام‌های جدید x_{1new} و x_{2new} در نظر بگیریم، دستور زیر می‌تواند برای دست‌یابی به برآوردهای این مثال با استفاده از روش شبه‌درست‌نمایی، مورد استفاده قرار گیرد:

```

model3=glm(ynew ~ x1new+x2new, family=quasi(variance=

```

```

Call:
glm(formula = combdat ~ logno2 * logtime, family = binomial
(link = logit), data = data)

Deviance Residuals:
    Min      1Q   Median      3Q     Max 
-1.9687 -0.5589  0.1499  0.4407  1.8066 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 0.18518   0.06248  2.964   0.003037 **  
logno2       1.03839   0.09096 11.416   < 2e-16 *** 
logtime      1.23738   0.09820 12.601   < 2e-16 *** 
logno2:logtime 0.22871   0.05971  3.830   0.000128 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 200.090 on 16 degrees of freedom
Residual deviance: 16.113 on 13 degrees of freedom
AIC: 109.66

Number of Fisher Scoring iterations: 3

```

شکل ۳.۲: نتایج برآش مدل ۱ با در نظر گرفتن اثرهای متقابل

"mu(1-mu)", link="logit"), start=c(0,1)).

۷.۲ تمرین‌ها

۱- در مطالعه‌ای بر روی موش‌ها، اثر دو نوع دارو را بر وجود تومور، مورد بررسی قرار داده‌ایم. نتایج آن در جدول زیر نمایش داده شده است:

۴.۲: تعداد موش‌های تحت درمان دو نوع دارو

داروی ۲	داروی ۱	
داشتن تومور		
نداشتن تومور		
۴	۵	
۱۲	۷۴	

با تعریف مناسب برای متغیر پاسخ Y و متغیر مستقل X ، مدل مناسبی برای داده‌های فوق معرفی کنید و به برآورد پارامترها بپردازید. نیکویی برآش مدل خود را چگونه ارزیابی می‌کنید؟ برآورد پارامترهای مدل خود را با استفاده از روش شبهدست‌نمایی نیز بیابید.

۲- در آزمایشی علاقه‌مند به بررسی میزان دوز دارو بر احتمال بهبودی

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

افراد هستیم. برای این منظور، دو مدل لوژستیک و پروبیت را در نظر می‌گیریم.

(آ) میزان دوزی از دارو را تعیین کنید که احتمال بهبودی 50% درصد را نتیجه می‌دهد. این میزان را تحت هر دو مدل لوژستیک و پروبیت بر اساس پارامترهای مدل به دست آورید.

(ب) اگر این مطالعه بر روی n بیمارانجام گرفته باشد و مدل لوژستیک زیر را برای آن در نظر گرفته باشیم، که در آن $Y_i = 1$ معادل پیشامد بهبودی فرد i است و x_i میزان دوز دارو برای فرد i است، نشان دهید لگاریتم نسبت درست‌نمایی برای آزمون $\theta = \alpha + \beta \sum x_i y_i$ رد می‌کند: (α و β معلوم فرض شده‌اند).

$$\Pr(Y_i = 1 | x_i) = (1 + \exp(-\alpha - \theta x_i))^{-1}, \quad i = 1, \dots, n$$

(۳) در مطالعه‌ای بر روی میزان رضایت شغلی افراد، متغیرهای مورد بررسی به صورت Y (متغیر پاسخ): رضایت‌مندی از شغل با دو سطح $0 = Y$ ناراضی، $1 = Y$ راضی؛ X : جنسیت با دو سطح $0 = X$ مرد، $1 = X$ زن؛ Z : سطح تحصیلات با سه سطح $1 = Z$ زیر دیپلم، $2 = Z$ دیپلم و فوق دیپلم، $3 = Z$ لیسانس و بالاتر در نظر گرفته شده‌اند. دو مدل لوژستیک به این داده‌ها برآش داده شده، که نتایج آن‌ها در جدول زیر خلاصه شده است:

۵.۲: نتایج حاصل از برآورد دو مدل بر اساس داده‌های رضایت شغلی

	برآورد (مدل ۱)	برآورد (مدل ۲)	پارامتر
$-0/5$	$-0/3$	مقدار ثابت	
$2/5$	$2/5$	$X = 1$	
$-$	$-0/02$	$Z = 1$	
$-$	$-0/1$	$Z = 2$	
$27/786$	$24/062$	منفی لگاریتم درست‌نمایی	

(آ) معادله‌ی رگرسیونی برآنده شده برای مدل ۱ را بنویسید و همه‌ی پارامترهای آن را تفسیر کنید.

(ب) اگر بخواهیم اثر متقابل جنسیت و سطح تحصیلات را بررسی کنیم، چه عبارتی را باید به مدل اضافه کنیم؟ آیا تفسیر پارامترها تغییر می‌کند؟ (به طور مختصر تشریح کنید).

(پ) فرض صفر «بی‌تأثیر بودن متغیر سطح تحصیلات» را با استفاده از روش نسبت

درست‌نمایی آزمون کنید.

ت) در حالتی که متغیر سطح تحصیلات را در برآش مدل در نظر نگیریم، جدول داده‌ها به صورت زیر خلاصه می‌شود:

جدول ۶.۲: وضعیت رضایت شغلی افراد به تفکیک جنس

وضع رضایت		جنس
جمع	راضی ناراضی	
۲۲	۷	مرد
۸۴	۷۹	زن
۱۰۶	۸۶	جمع
۱۵	۵	

براساس این جدول، مدل لوژستیک مناسب برای احتمال رضایت افراد را بنویسید و پارامترهای آن را به روش ماکسیمم درست‌نمایی براورد کنید. براوردهای مدل خود را با استفاده از روش شبه درست‌نمایی نیز بباید و نتایج خود را با براوردهای روش ماکسیمم درست‌نمایی مقایسه کنید. آیا مانده‌های پیرسون، مناسب بودن مدل را تأیید می‌کنند؟ همچنین مقدار نسبت بخت‌ها برای پیشامد رضایت افراد را محاسبه کنید و فاصله‌ی اطمینان ۹۵ درصدی برای آن را بباید.

ث) تابع درست‌نمایی مدل قسمت (ت) را بنویسید و مقدار لگاریتم درست‌نمایی را با جایگذاری براوردهای مدل از قسمت (ت) محاسبه کنید.

۴- برای جدول ۲۸.۱ از فصل ۱،

(آ) توزیع مناسب برای داده‌های جدول را ارائه کنید.

ب) مدل مناسب برای بررسی تأثیر دارو بر فراوانی مرگ در موش‌ها را بیان کنید. براساس پارامترهای این مدل، نسبت بخت‌ها را برای این جدول بیان کنید.

۵- فرض کنید $(n_i, \theta_i) \sim \text{bin}(y_i, \theta_i)$ باشد. نشان دهید که تابع درست‌نمایی $\theta_m, \dots, \theta_2, \theta_1$ به شرط y_1, \dots, y_m به صورت زیر است:

$$\ell(\theta_1, \dots, \theta_m | y) = K \prod_{i=1}^m \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i},$$

که در آن K مقدار ثابتی است که به پارامترها وابسته نیست. همچنین نشان دهید که تابع

فصل ۲. مطالعات مقطعی: داده‌های دارای پاسخ دودویی

درست‌نمایی لوجیت‌های $\alpha_1, \dots, \alpha_m$ (که $\alpha_i = \log \frac{\theta_i}{1-\theta_i}$) به صورت زیر است:

$$\ell(\alpha_1, \dots, \alpha_m | y) = K \exp\left\{\sum_{i=1}^m \alpha_i y_i\right\} \prod_{i=1}^m (1 + \exp(\alpha_i))^{-n_i}.$$

نشان دهید براوردگرهای ماکسیمم درست‌نمایی θ_i و α_i به ترتیب برابرند با $\hat{\theta}_i = \frac{y_i}{n_i}$ و

$$\hat{\alpha}_i = \log\left(\frac{y_i}{n_i - y_i}\right)$$

۶- برای یک جدول رده‌بندی شده‌ی $2 \times I$ ، مدل لوجیت زیر را در نظر گیرید:

$$\text{logit}[\Pr(Y = 1 | X = i)] = \alpha + \beta_i, \quad i = 1, 2, \dots, I$$

(آ) اگر $\beta_I = 0$ باشد، β_i را چگونه برای $i = 1, 2, \dots, I-1$ ، براورد می‌کنید. براورد خود را تفسیر کنید.

(ب)تابع درست‌نمایی را در صورت استقلال X و Y به دست آورید و α را براورد کنید.

(ج) اگر Y دو جمله‌ای و X متغیر تبیینی پیوسته باشد، چگونه می‌توان قبل از مدل‌بندی Y روی X ، دریافت که از چه نوع تابع ربطی استفاده کنیم؟

(د) داده‌های جدول زیر مربوط به ۱۳۰ درخت صنوبر است که در یک محیط مصنوعی برای مدت ۸ سال نگهداری شده‌اند.

جدول ۷.۲: وضعیت درخت‌های صنوبر بر اساس سطوح مختلف سولفور

سطح سولفور				تعداد آسیب دیده‌ها	تعداد کل
۱/۵۰	۱/۰۰	۰/۵۰	۰/۰۰		
۲۹	۲۰	۱۶	۲۵		
۳۸	۲۴	۲۷	۴۱		

جدول ۷.۲، تعداد درخت‌هایی را که آسیب دیده‌اند، به ازای سطوح مختلف سولفور نشان می‌دهد. اگر توزیع دو جمله‌ای را برای تعداد درختان آسیب دیده فرض کنیم و مدل لوجیت خطی

$$\text{logit}[\pi(x)] = a + bx$$

را برای $\pi(x)$ ببرازانیم که در آن $\pi(x)$ احتمال آسیب دیدن درخت در سطح x سولفور است، براوردهای پارامترها عبارت‌اند از $\hat{a} = 0/4719$ و $\hat{b} = 0/5465$.

آ) مقادیر برازش یافته تحت مدل را برای خانه‌های جدول بیابید.

ب) مانده‌های پیرسون را تحت مدل بیابید.

پ) مدلی را که فرض می‌کند غلظت سولفور تأثیری روی احتمال آسیب دیدن ندارد، برازش دهید. مقادیر برازانده شده و مانده‌های پیرسون آن‌ها را بیابید.

ت) کیش را بیابید و فرض $H_0 : b = 0$ را آزمون کنید.

۹- فرض کنید مدل خطی تعمیم‌یافته‌ی $\text{logit}[\pi(x)] = \alpha + \beta x$ با پاسخ دودویی Y مد نظر است و جدول زیر به دست آمده است:

جدول ۸.۲: مشاهدات مربوط به جدول پیش‌بینی حاصل از یک متغیر پاسخ و کمکی دو حالتی

n	$Y = 1$	x
۱۰۰	۵۰	۰
۱۰۰	۴۰	۱

براوردهای α و β را به دست آورید. $H_0 : \beta = 0$ را چگونه آزمون می‌کنید؟

۱۰- فرض کنید در یک مطالعه‌ی آینده‌نگر در صدد باشیم تأثیر داروی مختلف بر بیهوذی بیماری خاصی را بررسی کنیم. جدول زیر، داده‌های مربوط را نشان می‌دهد.

جدول ۹.۲: وضعیت بیماران بر اساس مصرف نوع دارو

	بیهوذی	عدم بیهوذی	
داروی ۱	۳۰	۷۰	۱۰۰
داروی ۲	۶۰	۴۰	۱۰۰
	۹۰	۱۱۰	۲۰۰

آ) با معرفی یک مدل و بررسی مدل پیشنهادی، چه آزمونی برای یافتن تأثیر داروی ۱ یا داروی ۲ بر وضعیت بیماران پیشنهاد می‌کنید؟

ب) براوردهای پارامترهای مدل ارائه کنید.

پ) یک فاصله‌ی اطمینان ۹۵٪ برای نسبت بخت‌ها بیابید و رابطه‌ی آن با پارامترهای مدل را مشخص کنید.