طرح‌های سه نقطه‌ای بهینه برای نمونه‌های کوچک در مدل لجستیک

زهرا منصوروار[۱] و هوشنگ طالبی[۲]

چکیده:

عموماً استنباط در یک مدل خطی تعمیم‌یافته بر اساس تقریب‌های مجانبی برای اریبی و ماتریس کوواریانس برآوردگر پارامترها انجام می‌گیرد. در آزمایش‌های با حجم نمونه کوچک، این تقریب‌ها ضعیف عمل می‌کنند زیرا برآوردگرهای اریب را نتیجه می‌دهند. بررسی طرح‌های بهینه در چنین آزمایش‌هایی در علوم زیستی و به خصوص در داروسازی به منظور تعیین سطح دوز دارو، از اهمیت ویژه‌ای برخوردار است. در این راستا، طالبی و راسل (۲۰۰۸) در مدل لجستیک، طرح‌های بهینه را برای نمونه‌های کوچک و با در نظر گرفتن معیار انتگرال میانگین مربعات خطای پیش بینی مورد بررسی قرار دادند. آن‌ها تحقیق خود را محدود به طرح‌های ۲—نقطه‌ای کردند. در این مقاله، به تعمیم طرح‌های ایشان به سه نقطه می‌پردازیم.

واژه‌های کلیدی: اریبی، انتگرال میانگین مربعات خطا، پیش بین خطی، مدل بتا—دو جمله‌ای، معیار بهینگی.

۱ مقدمه

طراحی یک آزمایش برای متغیر پاسخ دودوئی در علوم زیست شناسی به‌ویژه آزمایش‌های زیستی و نیز در علوم مهندسی به خصوص آزمون قابلیت اعتماد بسیار مهم است. به منظور مدل‌سازی متغیر پاسخ دودوئی از مدل‌های خطی تعمیم‌یافته استفاده می‌شود. مدل لجستیک در عمل کاربردی‌ترین مدل برای بررسی این گونه داده‌ها است که در خانواده مدل‌های خطی تعمیم‌یافته قرار می‌گیرد. بنابراین بررسی این مدل و ساختن طرح مناسب برای آن از اهمیت ویژه‌ای برخوردار است.

در حالت کلی یک طرح عبارت از تعیین تعداد و مقدار نقاط همراه با تعداد تکرار در هر نقطه است. در صورتی که برآورد پارامترهای مدل بر اساس معیاری دارای خواص بهینه باشد، به چنین طرحی، طرح بهینه گفته می‌شود. برای پیدا کردن طرح بهینه از معیارهای بهینگی استفاده می‌شود که این معیارها نشان دهنده خواسته‌های آزمایشگر از آزمایش مورد نظر هستند. معمولاً این معیارها تابعی از ماتریس اطلاع، به عنوان وارون ماتریس واریانس، هستند. در مدل‌های خطی تعمیم‌یافته برخلاف مدل‌های خطی، ماتریس اطلاع و در نتیجه معیارهای بهینگی به پارامترهای نامعلوم مدل بستگی دارد. اولین و شاید ساده‌ترین راه حل، در نظر گرفتن مقادیر اولیه‌ای برای پارامترهای نامعلوم مدل است. این رهیافت کلاسیک طرح‌های بهینه

[۱] گروه آمار، دانشگاه اصفهان
[۲] عضو هیئت علمی گروه آمار، دانشگاه اصفهان

موضعی را نتیجه می‌دهد که اولین بار توسط چرنف (۱۹۵۳) مطرح گردید. پس از آن پژوهشگران زیادی نظیر عبدالبسیط و پلاکت (۱۹۸۳) مین کین (۱۹۸۷) و فورد و همکاران (۱۹۹۲) به بررسی طرح‌های بهینه موضعی با معیارهای بهینگی گوناگونی پرداختند. یک تعمیم طبیعی برای طرح‌های بهینه موضعی، استفاده از یک توزیع پیشین برای پارامترهای نامعلوم مدل به جای یک حدس اولیه است. در این صورت طرح‌های مذکور را طرح‌های بهینه بیزی می‌نامند. محققین بسیاری نظیر سوتاکاوا (۱۹۷۲, ۱۹۸۰)، چالونر و لانتز (۱۹۸۹) و چالونر (۱۹۹۳) به بررسی طرح‌های بهینه بیزی پرداخته‌اند. یک رهیافت دیگر برای غلبه بر مشکل وابستگی معیارهای بهینگی به پارامترهای نامعلوم در مدل‌های خطی تعمیم‌یافته، روش مینی ماکس است. اولین بار سیتر (۱۹۹۲) این روش را برای مدل لجستیک به کار برد. روش مینی‌ماکس بدترین کارایی را در یک فضای از پیش تعیین شده برای پارامترها بهینه می‌کند. در حقیقت، این روش را می‌توان به عنوان روشی بین دو روش قبل در نظر گرفت.

معمولاً معیارهای بهینگی بر اساس ماتریس اطلاع و خواص مجانبی این ماتریس بنا شده‌اند. این خواص تا زمانی که حجم نمونه در آزمایش مورد نظر بزرگ باشد، برقرارند. بنابراین، در آزمایش‌هایی با حجم نمونه کوچک، بحث اریبی برآورد پارامترها مطرح می‌شود و در این صورت استفاده از معیارهای ذکر شده کفایت نمی‌کند. مسئله طرح‌های بهینه برای نمونه‌های کوچک توسط برخی از پژوهشگران مورد مطالعه قرار گرفته است، که از آن جمله می‌توان به سان و همکاران (۱۹۹۶)، سان و سوتاکاوا (۱۹۹۷)، راسل و همکاران (۲۰۰۹) و طالبی و راسل (۲۰۰۸) اشاره کرد.

طالبی و راسل برای یک آزمایش با پاسخ دودوئی و با احتمال موفقیت P در سطح دوز x، وجود عدم قطعیت را از طریق P فرض کردند. برای این منظور، $P \in (0, 1)$ را یک متغیر تصادفی و دارای توزیع بتا با میانگین π در نظر گرفتند. آن‌ها با در نظر گرفتن پیوند لجستیک بین π و سطح دوز x، مسئله طرح بهینه بیزی را برای طرح‌های ۲–نقطه‌ای با حجم نمونه کوچک بررسی کردند. آن‌ها با استفاده از برآوردگر بیز و با توجه به اریبی این برآوردگر، میانگین مربعات خطای $\hat{\pi}$ که در برگیرنده واریانس و اریبی است را به عنوان اساس معیار بهینگی خود در نظر گرفتند. سپس با مینیمم کردن انتگرال میانگین مربعات خطای برآوردگر بیز (BIMSE[۳])، به عنوان معیار بهینگی، طرح بهینه را به دست آوردند.

همان‌طور که اشاره شد، تحقیق بالا محدود به طرح‌های ۲–نقطه‌ای شده بود. ولی با توجه به نتایج به دست آمده توسط چالونر و لارنتز (۱۹۸۹) با افزایش عدم قطعیت، تعداد نقاط طرحی نیز افزایش می‌یابد. از سوی دیگر، عبدالبسیط و پلاکت (۱۹۸۳) نشان داده بودند که طرح‌های ۳–نقطه‌ای به مراتب استوارتر از طرح‌های ۲–نقطه‌ای عمل می‌کنند. در این تحقیق، ما با استفاده از معیار بهینگی طالبی و راسل (۲۰۰۸) ، BIMSE، روش آن‌ها را به طرح‌های ۳–نقطه‌ای تعمیم می‌دهیم. قابل ذکر است که تعمیم روش مذکور برای بیش از دو نقطه پیچیده است. در این مقاله با حل این مسئله در

[۳] Bayesian Integrated Mean Square Error

بخش ۲، طرح‌های ۳−نقطه‌ای را به دست می‌آوریم. هم‌چنین در بخش ۳، کارایی این طرح‌ها را با طرح‌های ۲−نقطه‌ای مقایسه و تحلیل خواهیم کرد.

۲ طرح‌های ۳−نقطه‌ای بر اساس معیار BIMSE

n_i آزمودنی را در هر سطح دوز x_i برای i برابر ۱، ۲ و ۳ در نظر بگیرید. فرض کنید u_{ij} برای j برابر ۱، ۲، ...، n_i، نشان دهنده j−امین پاسخ دودوئی در سطح دوز x_i است که مقادیر صفر و یک را به ترتیب به ازای پاسخ‌های منفی و مثبت به دوز مربوط اختصاص می‌دهد. متغیر تصادفی $y_i = \sum_{j=1}^{n_i} u_{ij}$ تعداد کل موفقیت‌های مشاهده شده در x_i را نشان می‌دهد.

حال متغیرهای تصادفی مستقل $P_i \in (0,1)$ را در نظر بگیرید که در واقع، احتمال یک موفقیت در سه سطح دوز x_i را نشان می‌دهد. مقدار P_i به سطح دوز x_i بستگی دارد، از این رو آن‌را با $P(x_i)$ نمایش می‌دهیم. مؤلفه‌های u_{ij} به شرط P_i از یکدیگر مستقل‌اند و نیز داریم

$$Pr(u_{ij} = 1|P_i) = P(x_i) = P_i.$$

هم چنین فرض کنید به ازای هر i، متغیرهای تصادفی مستقل P_i دارای توزیع پیشین بتا با پارامترهای a_i و b_i، $Beta(a_i, b_i)$، هستند که امید ریاضی و واریانس این متغیرهای تصادفی عبارتند از:

$$E(P_i) = \pi(x_i) = \pi_i = a_i/(a_i + b_i),$$

$$Var(P_i) = \rho_i \pi_i (1 - \pi_i), \qquad 0 < \rho_i < 1,$$

که در رابطه فوق $\rho_i = 1/(1 + a_i + b_i)$ است. از این رو، پارامترهای توزیع بتا یعنی a_i و b_i را می‌توان به صورت زیر نوشت

$$a_i = \frac{\pi_i}{\tau_i}, \qquad b_i = \frac{1 - \pi_i}{\tau_i}, \qquad (1)$$

که در آن $\tau_i = 1/(a_i + b_i) = \rho_i/(1 - \rho_i)$ یک اندازه پراکندگی پیشین است.

در هر سطح دوز x_i، متغیر تصادفی y_i به شرط $P(x_i)$ دارای توزیع دو جمله‌ای با پارامترهای n_i و P_i، $Bin(n_i, P_i)$، است. از این‌رو، احتمال موفقیت $P(x_i)$ دارای توزیع پسین بتا با پارامترهای $y_i + a_i$ و $n_i - y_i + b_i$ خواهد بود. لذا میانگین پسین μ_i و واریانس پسین σ_i^2 به ترتیب به صورت زیر محاسبه می‌شوند

$$\mu_i = \frac{y_i + a_i}{n_i + (a_i + b_i)};$$
$$\sigma_i^2 = \mu_i (1 - \mu_i) \Big[\frac{\tau_i}{1 + \tau_i (1 + n_i)} \Big] \qquad (2)$$

توزیع حاشیه‌ای برای هر یک از متغیرهای تصادفی مستقل y_i نیز از یک توزیع بتا−دوجمله‌ای با پارامترهای n_i، π_i و τ_i، $BetaBin(n_i, \pi_i, \tau_i)$، پیروی می‌کند که تابع چگالی آن عبارت است از

$$f(y_i, n_i; \pi_i, \tau_i) = \qquad (3)$$
$$\binom{n_i}{y_i} \frac{B(y_i + a_i, n_i - y_i + b_i)}{B(a_i, b_i)}, \quad y_i = 0, 1, ..., n_i$$

میانگین پسین $P(x_i)$ یعنی μ_i که در رابطه (۲) محاسبه گردید، در حقیقت برآوردگر بیز π_i بنابر تابع زیان مربع خطا است، که این برآوردگر را با $\hat{\pi}_i$ نمایش می‌دهیم.

بدون آن که کلیت مسئله را از دست دهیم و به منظور سهولت در محاسبات، توجه خود را معطوف به طرح‌های متقارن ۳−نقطه‌ای خواهیم کرد. یعنی نقاط را متقارن

حول صفر با فواصل یکسان d در نظر می‌گیریم و نیز تعداد مشاهدات در هر نقطه را برابر با n فرض می‌کنیم. لذا نقاط طرح عبارتند از

$$x_1 = -d, \quad x_2 = 0, \quad x_3 = d = -x_1 \qquad (4)$$

البته از آنجا که طالبی و راسل (۲۰۰۸) دریافتند با فرض برابری n_iها و هم‌چنین r_iها، طرح‌هایی که به دست می‌آیند متقارن هستند، از این رو متقارن فرض کردن طرح، غیرمنطقی نیست.

حال فرض کنید، پیوند میان میانگین پیشین π و x از مدل لجستیک زیر پیروی می‌کند

$$\pi(x) = \pi = 1/\left[1 + exp(-\eta_x)\right] \qquad (5)$$

که در آن $\eta_x = \beta_0 + \beta_1 x$، پیش بین خطی و β_0 و β_1 پارامترهای مجهول مدل هستند. لذا اگر η_i نشان دهنده پیش بین خطی در سطح دوز x_i باشد، می‌توان آن‌را به صورت زیر نوشت

$$\eta_i = \beta_0 + \beta_1 x_i = ln\Big(\frac{\pi_i}{1-\pi_i}\Big), \quad i = 1,2,3 \quad (6)$$

اما با جایگذاری $\hat{\pi}_i$ به جای π_i در عبارت فوق، معادلات (۶) را به صورت زیر می‌توان نوشت

$$z_i = -A_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1,2,3$$

که در آن $A_i = \ln(\frac{n_i - y_i + b_i}{y_i + a_i})$ و $-A_i$ را با z_i نمایش می‌دهیم. واضح است، به دلیل وجود سه معادله و در نتیجه سه نقطه در فضای دو بعدی، مقادیر ضرایب خط یعنی پارامترهای β_0 و β_1 را نمی‌توان از حل این معادلات به دست آورد. لذا به منظور غلبه بر مشکل مذکور و یافتن برآوردگرهایی یکتا برای β_0 و β_1، بهترین خط رگرسیون را برای سه نقطه x_1، x_2 و x_3 برازش می‌دهیم و از این طریق ضرایب β_0 و β_1 را برآورد می‌کنیم. از این رو، برآوردگرهای β_1 و β_0 عبارتند از

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{3}(x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^{3}(x_i - \bar{x})^2}, \qquad (7)$$

$$\hat{\beta}_0 = \bar{z} - \hat{\beta}_1 \bar{x}. \qquad (8)$$

بنابر نقاط انتخاب شده در رابطه (۴)، $\bar{x} = 0$ خواهد بود و نیز از آن جا که $z_i = -A_i$ تعریف شده است، لذا $\bar{z}$ به صورت زیر محاسبه می‌شود

$$\bar{z} = \frac{1}{3}\sum_{i=1}^{3} z_i = -\frac{1}{3}(A_1 + A_2 + A_3).$$

از این رو، با جایگذاری مقادیر x_i، $\bar{x}$، z_i و $\bar{z}$ در روابط (۷) و (۸)، برآوردگرهای $\hat{\beta}_1$ و $\hat{\beta}_0$ را می‌توان به صورت زیر ساده نمود

$$\hat{\beta}_1 = \frac{A_1 - A_3}{2d},$$

$$\hat{\beta}_0 = \bar{z} = -\frac{1}{3}(A_1 + A_2 + A_3)$$

لذا در سطح دوز x، برآورد پیش بین خطی یعنی $\hat{\eta}_x$ عبارت است از

$$\hat{\eta}_x = \hat{\beta}_0 + \hat{\beta}_1 x = \frac{-2d(A_1 + A_2 + A_3) + 3(A_1 - A_3)x}{6d}$$

از این رو، بنابر رابطه فوق و تحت مدل مورد اشاره در (۵)، برآوردگر میانگین احتمال موفقیت $\pi(x)$ تحت مدل لجستیک، به صورت زیر محاسبه می‌شود

$$\hat{\pi}(x) = \qquad (9)$$
$$\left[1 + \left\{(B_1)^{2d-3x}(B_2)^{2d}(B_3)^{2d+3x}\right\}^{1/6d}\right]^{-1}$$

که در آن برای هر i، $B_i = (n_i - y_i + b_i)/(y_i + a_i)$.
حال با توجه به نقاط در (۴) و از آن جا که $2d = x_3 - x_1$ است، می‌توان $\hat{\pi}$ محاسبه شده در (۹) را با فرض $\beta_0 = 0$ بدون آن که کلیت را از دست بدهیم، به صورت زیر نوشت

$$\hat{\pi}(\eta) =$$
$$[1+\{(B_1)^{\eta_3-\eta_1-3\eta}(B_2)^{\eta_3-\eta_1}(B_3)^{\eta_3-\eta_1+3\eta}\}^{\frac{1}{3(\eta_3-\eta_1)}}]^{-1}$$

با جایگزین کردن (۵) در عبارات (۱) داریم

$$a_i = [\tau_i\{1 + exp(-\eta_i)\}]^{-1},$$
$$b_i = \tau_i^{-1}[1 - 1/\{1 + exp(-\eta_i)\}],$$

هم چنین با جایگزین کردن روابط فوق به جای a_i و b_i در B_i، می‌توان $\hat{\pi}(\eta)$ را به عنوان تابعی از η، η_1 و η_3 بیان نمود که تنها از طریق τ_i به توزیع پیشین بستگی دارد (واضح است که تحت شرایط ذکر شده، $\eta_2 = 0$ است).
حال $MSE(\hat{\pi})$ نیز به صورت زیر محاسبه می‌شود

$$MSE(\hat{\pi}) = \sum_{y_1=0}^{n_1}\sum_{y_2=0}^{n_2}\sum_{y_3=0}^{n_3}(\hat{\pi} - \pi)^2 \times$$
$$Pr(y_1 = \mathbf{y_1})Pr(y_2 = \mathbf{y_2})Pr(y_3 = \mathbf{y_3})$$

که در آن متغیرهای تصادفی مستقل y_1، y_2 و y_3 هر یک دارای توزیع بتا—دوجمله‌ای با تابع چگالی ارائه شده در (۳) هستند.
در مدل لجستیک معرفی شده در (۵)، هدف برآورد پارامترهای β_0 و β_1 و در نتیجه برآورد η به منظور پیش بینی احتمال موفقیت $\pi(\eta) = 1/[1 + exp(-\eta)]$ است.
از این رو، $MSE(\hat{\pi})$ به عنوان اساس معیار بهینگی در نظر گرفته می‌شود. لذا مینیمم کردن انتگرال میانگین مربعات خطای $\hat{\pi}$ به عنوان معیار بهینگی برای انتخاب طرح معرفی می‌شود که به صورت زیر تعریف می‌گردد (طالبی و راسل، ۲۰۰۸)

$$BIMSE = \int_{-\infty}^{\infty} MSE(\hat{\pi}(\eta))d\eta.$$

البته لازم به ذکر است، در غیاب یک عبارت صریح برای $MSE(\eta)$ بر حسب η، $BIMSE$ توسط انتگرالگیری $MSE(\eta)$ از $\eta_{0.0001}$ تا $\eta_{0.9999}$ تقریب زده می‌شود که در آن η_q در رابطه زیر صدق می‌کند

$$q = 1/[1 + exp(-\eta_q)], \qquad 0 < q < 1.$$

هم چنین، ذکر این نکته ضروری است که در این جا، محاسبه انتگرال مذکور به طور عددی و توسط قاعده سیمپسون و با تقسیم فاصله $\eta_{0.0001}$ تا $\eta_{0.9999}$ به ۱۰۰ زیر بازه انجام گرفته است.
با معلوم بودن مقادیر n_i و τ_i، برای هر i، مسئله طرح عبارت است از یافتن مقادیر η_1 و η_3 به گونه‌ای که $BIMSE$ مینیمم شود. ما بدون از دست دادن کلیت، هر یک از n_iها را برابر با n و نیز هر یک از τ_iها را برابر با τ فرض کردیم و با در نظر گرفتن مقادیر متفاوتی برای n و τ و نیز با استفاده از روش‌های بهینه سازی عددی در نرم افزار R، مقادیر η_1 و η_3 و نیز مقادیر مینیمم $BIMSE$ را به دست آوردیم. این مقادیر در جدول (۱) ارائه شده‌اند.
البته باید توجه شود، از آن جا که نقاط x_1 و x_3 متقارن هستند، لذا η_1 و η_3 نیز متقارن خواهند بود.
با توجه به جدول ۱ می‌توان دریافت به ازای τ ثابت، با افزایش n، $BIMSE$ افزایش می‌یابد ولی مقادیر η_3 (d) کاهش می‌یابند. هم چنین هر چه τ کوچک‌تر باشد، میزان

تغییرات در مقادیر $\eta_{۳}$ کمتر است.

۳ نتیجه‌گیری

با مقایسه طرح‌های به دست آمده در جدول ۱ برای طرح‌های ۳–نقطه‌ای با طرح‌های ۲–نقطه‌ای محاسبه شده توسط طالبی و راسل (۲۰۰۸) می‌توان نتیجه گرفت، حدود نقاط طرحی برای طرح‌های ۳–نقطه‌ای بیشتر از طرح‌های ۲–نقطه‌ای است ولی با کاهش τ، این حدود به طرح‌های ۲–نقطه‌ای نزدیک می‌شوند.

با مقایسه مقادیر $BIMSE$ برای طرح‌های ۲–نقطه‌ای و ۳–نقطه‌ای نیز در می‌یابیم، این مقادیر در طرح‌های ۳–نقطه‌ای کمتر از طرح‌های ۲–نقطه‌ای هستند. لذا می‌توان نتیجه گرفت که طرح‌های ۳–نقطه‌ای در مقابل تغییرات τ پایدارترند و این موضوع نظر چالونر و لارنتز (۱۹۸۹) را تأیید می‌کند که با افزایش عدم قطعیت، نقاط طرحی نیز افزایش می‌یابند تا طرح‌های پایدارتری ارائه نمایند.

هم چنین، در صورتی که روش پیشنهاد شده را با فرض $n_{۲} = ۰$ برای ۲ نقطه به کار گیریم، به همان نتایج به دست آمده توسط طالبی و راسل (۲۰۰۸) دست می‌یابیم که این مطلب می‌تواند صحت روشی که به‌کار گرفتیم را به اثبات برساند.

مراجع

[1] bdelbasit, K.M. and Plackett, R.L. (1983). Experimental Design for binary data, *Journal of the American Statistical Association*, 78, 90-98.

[2] haloner, K. (1993). A Note on Optimal Bayesian Design for Nonlinear Problems, *Journal of Statistical Planning and Inference*. 37, 229-235.

[3] haloner, K. and Larntz, K. (1989). Optimal Bayesian Experimental Design Applied to Logistic Regression Experiments, *Journal of Statistical Planning and Inference* , 21, 191-208.

[4] hernoff, H. (1953). Local Optimal Designs for Estimating Parameters, *Annals of Mathematical Statistics*, **24**, 586-602.

[5] ord, I.; Torsney, B. and Wu, C.F.J. (1992). The Use of a Canonical Form in The Construction of Locally Optimal Designs for Nonlinear Problems, *Journal of the Royal Statistical Society, Series B*, 54, 569-583.

[6] inkin, S (1987). Optimal Design for Binary Data, *Journal of the American Statistical Association*, **82**, 1098-1103.

[7] ussell, K.G.; Eccelston, J.A., Lewise, S.M. and Woods, D.C. (2009). Design Considerations for Small Experiments and Simple Logistic Regression, *Journal of Statistical Computation and Simulation*, 79, 81-91.

[8] itter, R.R. (1992). Robust Designs for Binary Data, *Biometrics*, 48, 1145-1155.

[9] un, D. and Tsutakawa, R.K. (1997). Bayesian Design for Dose-Response Curves with Penalized Risk, *Biometrics*, 53, 1262-1273.

[10] un, D., Tsutakawa, R.K. and Lu, W. (1996). Bayesian Design of Experiment for Quantal Responses : What's Promised Versus What's Delivered, *Journal of Statistical Planning and Inference*. 52, 289-306.

[11] alebi, H. and Russell, K.G. (2008). Optimal Designs for Simple Logistic models with Success Probability Uncertainty for Small Experiments, *Australian and New zeland Journal of Statistics* (To be appear).

[12] sutakawa, R.K. (1972). Design of an Experiment for Bioassay, *Journal of the American Statistical Association*, 67, 584-590.

[13] sutakawa, R.K. (1980). Selection of Dose Levels for Estimating a Percentage Point of a Logistic Quantal Response Curve, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 29, 25-33.