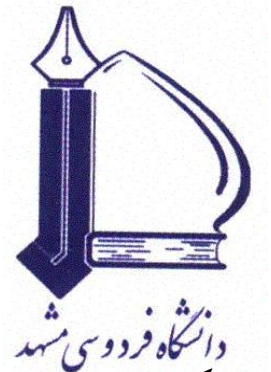

دانشکده علوم ریاضی

پایان‌نامه جهت اخذ درجه کارشناسی ارشد

آمار ریاضی

طرح آزمایش های بهینه برای مدل رگرسیون پواسون

استاد راهنما

دکتر مجید سرمد

استاد مشاور

دکتر غلامحسین شاهکار

نگارش

منصوره روشن نژاد

شهریور ۱۳۸۸

قدردانی

اینک که در پرتو مهر ایزدی و در جوار بارگاه ملکوتی امام رئوف حضرت رضا (ع) به پایان مرحله‌ای دیگر از دوران تحصیلم رسیده‌ام، بر خود لازم می‌دانم از عزیزانی که اینجانب را طی این مسیر یاری نموده‌اند، تشکر و قدردانی نمایم.

در ابتدا از تمامی اساتید ارجمندم در دانشگاه فردوسی مشهد به ویژه جناب آقای دکتر ارقامی که با صبر و دلسوزی مرا راهنمایی نمودند، صمیمانه تشکر می‌نمایم. همچنین مایلم تشکر خالصانه خود را از جناب آقای دکتر سرمد استادیار محترم گروه آمار دانشگاه فردوسی مشهد که راهنمایی این رساله را بر عهده داشتند و مرا در دانش خود سهیم نمودند، ابراز نمایم. همچنین از آقای دکتر شاهکار که سمت مشاور این رساله را بر عهده داشتند تشکر می نمایم.همین طور از آقای دکتر طالبی و آقای پورسینا به دلیل وقتی که در اختیارم گذاشتند تشکر می نمایم. و مایلم مراتب تشکر و قدردانی خود را از آقایان دکتر هادی جباری نوقابی و دکتر رزمخواه که داوری این رساله را بر عهده داشتند و اینجانب را از راهنمایی‌های خود بهره‌مند ساختند اعلام نمایم. لازم است از منشی محترم گروه آمار سرکار خانم سلیمانی و مسئولین کتابخانه آقای داوودنژاد و سرکار خانم صادقی به دلیل کمک صمیمانه و تهیه مراجع مورد نیاز تشکر ویژه می نمایم. همچنین از آقای وطن دوست و خانم نخعی مسئول رایانه دانشکده تشکر و قدردانی می‌نمایم. مایلم از آقای شاهینی و آقای رمضانپور که راهنمایی و مساعدت‌های لازم را به منظور تایپ رساله در محیط تک میکر داشته‌اند و سایر عزیزانی که به نحوی در تهیه این رساله مرا کمک نمودند تشکر نمایم.

از پدر و مادر عزیزم، همسر گرانقدرم و مادر مهربان همسرم که همواره مشوق و حامی من بوده‌اند و با صبر و بردباری بی نظیرشان مرا در طی این مسیر یاری نموده‌اند، خالصانه سپاسگزاری می‌نمایم.

چکیده

در سال هاي اخير، توجه به طرح آزمايش هاي بهينه در مدل هاي خطي تعميم يافته افزايش يافته است. با وجود اين، بيشتر تحقيقات جاري روي مدل داده هاي دودويي، مخصوصاً مدل رگرسيون لجستيك يك متغيره مرتبه اول متمركز است. پژوهش حاضر، اين موضوع را روي مدل داده هاي شمارشی تعميم داده است. هدف اصلي اين تحقيق توسعه و بسط طرح آزمايش هاي كارا و استوار روي مدل رگرسيون پواسون در مطالعات سم‌شناسی است.

طرح‌های D- بهينه برای مدل يک سمی مرتبه اول و مدل دو سمی با اثرمتقابل بررسی شده و وابستگی آنها به پارامترهای مدل ارزيابی شده است. کاربرد طرح های D- بهينه، به دليل اين که طرح های بهينه برحسب سطوح دوز مؤثر (ED) به پارامترهای مجهول بستگی دارند، محدود است. بنابراين، برخی از اين طرح های عملی مانند طرح‌های فضای برابر و طرح‌های D- بهينه شرطی که برحسب سطوح دوز مؤثر بوده و مستقل از پارامترها می‌باشند، بررسی شده‌اند. اين طرح های عملی وقتی که فضای طرح محدود باشد ، کاملاً کارا هستند.

طرح های تهيه شده برحسب سطوح دوز مؤثر شبيه طرح‌هایD - بهينه در برابر عدم اطلاع از پارامترها استوار نيستند. برای بررسی اين مسئله، طرح های دنباله‌ای برای مدل‌های رگرسيون پواسون پيشنهاد شدند. طرح‌های کاملاً دنباله‌ای مورد بررسی قرار گرفتند و به گونه‌ای تهيه شدند که نسبت به عدم اطلاع از پارامترها، کارا و استوار باشند.

واژه‌های کلیدی: *طرح، رگرسيون پواسون، مدل‌های خطی تعميم‌يافته*

پیش‌گفتار

در بسیاری از موارد به خصوص در علوم مهندسی متغیر پاسخ متغیری پیوسته است. معمولاً بر روی این متغیرها مدل‌های خطی به خوبی قابل برازش است. برای چنین مدل‌هایی طرح‌های گوناگونی تاکنون ارائه شده‌اند که از آن جمله می‌توان به طرح‌های کسری اشاره نمود. طرح‌های دیگری نیز در صنعت و علوم داروشناسی به صورت وسیعی استفاده می‌شوند که فرضیات اولیه‌ی طرح‌های معمول برای آن‌ها برقرار نیست. یکی از مهم‌ترین مثال‌ها زمانی است که جواب نهایی یک متغیر شمارشی است. این گونه متغیرها به‌وسیله مدل‌های خطی قابل بیان نیستند و برای بررسی آن‌ها نیاز به مدل‌های خطی تعمیم‌یافته می‌باشد.

با توجه به اهمیت استفاده از مدل‌های خطی تعمیم‌یافته به خصوص در علوم داروشناسی، در فصل دوم به معرفی مدل‌های خطی تعمیم‌یافته، ساختار آن‌ها و روش به‌دست آوردن برآورد پارامترها و ماتریس اطلاع در این مدل‌ها پرداخته‌ایم. سپس مدل پواسون را که در عمل از کاربرد بیشتری در داده‌های شمارشی برخوردار است، مورد بررسی قرار داده ایم.

در فصل سوم طرح‌های D- بهینه برای مدل یک سمی مرتبه اول و مدل دو سمی با اثرمتقابل را بررسی کرده ایم و سپس به برخی از طرح‌های عملی برای این دو مدل مانند طرح فضای برابر و طرح D- بهینه شرطی پرداخته‌ایم. در نهایت استواری این طرح‌ها را نسبت به بدمشخص‌سازی پارامترها بررسی کردیم.

در فصل چهارم نظریه‌ی هم ارزی کلی را شرح داده ایم و از آن برای تحقیق بهینگی طرح‌های بهینه موضعی در فصل سوم بهره گرفته ایم. این کار لازم است، زیرا برای

به‌دست آوردن طرح‌های بهینه از روش‌های عددی استفاده شده است و به‌طور تحلیلی آن‌ها را به‌دست نیاورده ایم، لذا بهینگی آن‌ها زیر سؤال است.

راه‌حل ارائه شده در فصل سوم جایگزینی پارامترهای مجهول مدل با یک مقدار اولیه‌ی خوب است. این راه‌حل به طرح‌های بهینه موضعی منجر می‌شود. در ابتدای فصل پنجم به بررسی راه‌حل ارائه شده در فصل سوم می‌پردازیم. ایرادهای این روش به صورت عمده عبارتند از:

۱. عدم وجود مقدار اولیه‌ی خوب برای پارمترها

۲. عدم استواری این طرح‌ها در انتخاب های ناصحیح پارامترها

با توجه به ایرادهای ذکر شده، در فصل پنجم راه‌حلی برای غلبه بر مشکل وابستگی معیارهای بهینگی مطرح می‌شود. در این فصل ابتدا به معرفی طرح‌های بهینه دنباله‌ای می‌پردازیم. سپس این طرح‌ها را برای مدل پواسون توضیح می دهیم. با ارائه الگوریتمی و با روش‌های تکراری، طرح بهینه دنباله‌ای را به‌دست می‌آوریم.(برنامه این الگوریتم در ضمیمه پایان نامه موجود می‌باشد).

فهرست مطالب

لیست تصاویر

لیست جداول

فصل ۱

تعاریف و مقدمات

۱.۱ مقدمه

در این فصل به اهمیت موضوع و مرور پیشینه و انگیزه تحقیق حاضر می‌پردازیم. سپس مروری بر طرح‌های بهینه در مدل‌های خطی تعمیم‌یافته خواهیم داشت. مراجع [۲]، [۴۰]و [۵۶] از منابع اصلی این فصل هستند.

۲.۱ نمادگذاری

ابتدا معرفی خلاصه‌ای از نمادهای مورد استفاده را خواهیم داشت.

متغیر پاسخ را با Y نشان می‌دهیم و یک مجموعه از متغیرهای توضیحی به وسیله $x_p, \ldots, x_۲, x_۱$ نشان داده می‌شود. معادله پیوند به صورت زیر می‌باشد:

$$g(E(Y)) = \beta_۰ + \beta_۱ x_۱ + \ldots + \beta_p x_p.$$

برای متغیرهای پاسخ $Y_n, \ldots, Y_۲, Y_۱$، معادله پیوند به شکل ماتریسی زیر در می‌آید:

$$g(E(\mathbf{Y})) = \mathbf{X}\beta,$$

که در آن

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

بردار متغیرهای پاسخ است.

$$g[E(\mathbf{Y})] = \begin{bmatrix} g[E(Y_1)] \\ \vdots \\ g[E(Y_n)] \end{bmatrix} \qquad (۱.۱)$$

نشان‌دهنده‌ی بردار توابع $E(Y_i)$ می‌باشد که تابع g برای تمام عناصربردار یکسان است و

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

بردار پارامترها است و X ماتریسی است که عناصر آن نشان‌دهنده‌ی سطح متغیر توضیحی رسته‌ای و یا نمایانگر مقدار متغیر توضیحی پیوسته است.

برای یک متغیرتوضیحی x (مانند وزن) مدل شامل جمله‌ی βx است که پارامتر β نشان‌دهنده‌ی تغییر در متغیر پاسخ به ازای تغییر یک واحدی در متغیر x است.

برای متغیر توضیحی رسته‌ای، پارامترهایی وجود دارد که سطوح مختلف یک عامل را نشان می‌دهند.

اگر تعداد پارامترهای مدل p باشد و n مشاهده داشته باشیم، آن‌گاه y بردار تصادفی $n \times 1$ می‌باشد، β بردار $p \times 1$ از پارامترها است و X یک ماتریس $n \times p$ از ثابت های معلوم می‌باشد. X اغلب **ماتریس طرح** و $\mathbf{X}\beta$ **مؤلفه‌ی خطی مدل** نامیده می‌شود.

تعریف ۱.۲.۱. تابعی که $E(Y_i) = \mu_i$ را به مؤلفه‌ی خطی $\mathbf{x_i'}\beta$ مربوط می‌سازد **تابع پیوند** نامیده می‌شود. مانند

$$g(\mu_i) = \mathbf{x_i'}\beta \qquad (۲.۱)$$

که در آن g تابع پیوند نامیده می‌شود.

مدل‌های خطی زیر را در نظر بگیرید:

$$E(Y_i) = \mu_i = \mathbf{x_i'}\beta \quad ; \quad Y_i \sim N(\mu_i, \sigma^2) \qquad (۳.۱)$$

که در آن $Y_n, \ldots, Y_1$ متغیرهای تصادفی مستقل می‌باشد و بردار $\mathbf{x}_i'$ نشان‌دهنده‌ی iامین سطر ماتریس طرح $\mathbf{X}$ است. این مدل معمولاً به صورت ماتریسی زیر نوشته می‌شود:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}, \qquad (۴.۱)$$

که در آن

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad , \mathbf{X} = \begin{bmatrix} \mathbf{x_1'} \\ \vdots \\ \mathbf{x_n'} \end{bmatrix} \quad , \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad , \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

و e_iها متغیرهای تصادفی مستقل و هم‌توزیع با $e_i \sim N(0, \sigma^2)$ ، $i = 1 \ldots n$ می‌باشند.

مثالی در این مورد می‌تواند رابطه میان وزن زمان تولد نوزاد و سن حاملگی مادر باشد. اکنون حالتی را در نظر بگیرید که Y_iها دارای توزیع نرمال نباشند و یا رابطه میان متغیرهای پاسخ و متغیرهای توضیحی به صورت مدل (۳.۱) نباشد. مثلاً بررسی مرگ‌ومیر ماهیان (متغیر پاسخ) در مقابل استفاده‌ی مقدار مشخصی از سم (متغیر توضیحی)، در این مثال متغیر پاسخ پیوسته نیست و یک متغیر دوجمله‌ای است. برای بررسی چنین مسائلی نیاز به مدل‌های دیگری با فرضیاتی به غیر از فرضیات مدل خطی (۳.۱) داریم. در این مثال نیاز به مدلی داریم که توزیع متغیرهای پاسخ متعلق به توزیع دوجمله‌ای باشد و علاوه بر این بنا به تجربه می‌دانیم که **تابعی غیرخطی** $E(Y_i) = \mu_i$ را به $\mathbf{x}_i'\beta$ مربوط می‌سازد.از جمله‌ی این مدل‌ها می‌توان به مدل دوجمله‌ای که در آن $g(\mu_i) = \ln \frac{p_i}{1-p_i}$ و مدل پواسون که در آن $g(\mu_i) = \ln(\mu_i)$ می‌باشد، اشاره نمود. یکی از ویژگی‌های مدل‌های خطی تعمیم یافته تابع پیوند غیر خطی است، تعریف دقیق این مدل‌ها را تا مروری بر پیشینه موضوع به تعویق می‌اندازیم.

تعریف ۲.۲.۱. به‌طورکلی، دوز مؤثر یا به اختصار ED، به عنوان دوز یا مقدار سمی تعریف می‌شود که منجر به یک نسبت مشخص از مرگ و میر یا بقاء می‌شود. به عنوان مثال ED_{30} سم A دوزی از سم A است که منجر به۷۰% مرگ ومیر می‌شود. برای آزمایشات یک سمی، این تعریف، صریح و روشن است. برای آزمایشات چند سمی، نیاز به تشخیص اثر ترکیبی تمام سموم و ترکیب هر سم به تنهایی داریم. برای این منظور **دوز مؤثر مدل** یا به اختصار MED را تعریف می‌کنیم و به عنوان تمام ترکیبات ممکن دوزهای تمام سموم که منجر به یک مرگ ومیر مشخص یا بقای

مشخص می‌شوند، در نظر می گیریم. به عنوان نمونه، در یک آزمایش دو سمی، $MED_{۳۰}$، تمام ترکیبات ممکن دوزهای سمی است که مرگ ومیر ۷۰٪ را ایجاد می کنند. هرچند، دوز مؤثر انفرادی یا به اختصار IED یک سم، دوزی از سم است که منجر به مرگ ومیر یا بقای مشخص است. روشن است برای آزمایشات تک سمی MED و IED یکسان هستند.

۳.۱ پیشینه موضوع و انگیزه تحقیق

محققی را در نظر بگیرید که بر روی داروی ضد سرطان مطالعه و بررسی می‌کند و بنا به تجربیات قبلی می‌داند که تعداد کلونی‌های سرطانی، معمولاً یک متغیر پواسون در نظر گرفته می‌شود. برای استفاده‌ی صحیح از داروی مورد نظر می‌خواهد مدل رگرسیون پواسون را که رابطه‌ی میان تعداد کلونی‌ها و غلظت یا دوز دارو را توصیف می‌کند، مشخص نماید. با توجه به مدل (۳.۱)، واضح است که مسأله در فرضیات مدل‌های خطی صدق نمی‌کند و یکی از مسائل مدل‌های خطی تعمیم‌یافته است. از این نوع مسائل در داروشناسی و سم‌شناسی فراوان است.

مدل‌های خطی تعمیم‌یافته برای داده‌های دودویی و شمارشی، در کتاب‌ها و مقالات بسیاری بررسی شده‌اند [۴۰]. اخیراً، طرح آزمایش‌ها برای مدل داده‌های دودویی، مخصوصاً برای مدل رگرسیون لجستیک، که کاربرد فراوانی در علوم زیست شناسی و دارو شناسی دارد، توجه زیادی را به خود معطوف ساخته و بسیاری از کتاب‌ها و مقالات، به این مبحث پرداخته اند. با وجود این، کار کمی بر روی طرح آزمایش‌های کارا برای مدل **داده‌های شمارشی** مانند مدل رگرسیون پواسون انجام شده است. در این تحقیق، سعی داریم این شکاف را به وسیله توسعه طرح آزمایش‌های کارا و استوار برای مدل رگرسیون پواسون پرکنیم.

یک بررسی سم‌شناسی شامل اجرای غلظت‌های مختلف سموم بر روی ارگانیزم‌ها و اندازه گیری اثر شاخص‌هایی همچون مرگ ومیر و رشد و صدمه به تولد و تناسل و... را در بر می‌گیرد. (معمولاً این پاسخ‌ها نمی‌تواند به‌طور مناسبی به عنوان متغیرهای

تصادفی نرمال مدل‌بندی شود). شاخص‌هایی همچون مرگ‌و میر، پاسخ‌های دودویی و شاخص‌هایی همچون صدمه به تولد وتناسل پاسخ‌هایی شمارا دارند. در چنین موقعیت‌هایی، مدل‌های خطی کلاسیک عملی نیست و معمولاً **مدل‌های خطی تعمیم‌یافته** برای تحلیل به کارگرفته می‌شود. مثال‌های مشابه فراوانی در کتاب‌ها وجود دارد [۳۵].

۴.۱ مروری بر طرح در مدل‌های خطی

در این بخش به معرفی طرح‌ها در مدل‌های خطی می‌پردازیم تا خواننده بتواند ارتباط مؤثری بین این طرح‌ها و طریقه به‌دست آمدن آن‌ها و تفاوت آن‌ها با مدل‌های خطی تعمیم‌یافته برقرار کند. در این قسمت مدل‌هایی به شکل زیر را در نظر می گیریم :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}, \qquad (۵.۱)$$

که در آن

$$\mathbf{y} = \begin{bmatrix} y_۱ \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} ۱ & x_{۱۱} & x_{۱۲} & \cdots & x_{۱p} \\ ۱ & x_{۲۱} & x_{۲۲} & \cdots & x_{۲p} \\ \vdots & \vdots & \vdots & & \vdots \\ ۱ & x_{n۱} & x_{n۲} & \cdots & x_{np} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_۰ \\ \vdots \\ \beta_p \end{bmatrix}, \qquad \mathbf{e} = \begin{bmatrix} e_۱ \\ \vdots \\ e_n \end{bmatrix},$$

که e_iها متغیرهای تصادفی مستقل و با توزیع $e_i \sim N(۰, \sigma^۲)$,$i = ۱, \ldots, n$ می‌باشند. از رگرسیون می‌دانیم که برآوردگر درستنمایی ماکسیمم β به صورت زیر به‌دست می‌آید:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'\mathbf{y}. \qquad (۶.۱)$$

از آنجا که برآوردهای به‌دست آمده نااریب هستند، رابطه زیر میان ماتریس واریانس-کوواریانس و ماتریس اطلاع برقرار است:

$$cov(\hat{\beta}) = \sigma^۲(\mathbf{X}'\mathbf{X})^{-۱} = \mathbf{I}^{-۱}(X, \beta). \qquad (۷.۱)$$

که در آن $\mathbf{I}$ ماتریس اطلاع است.

۵.۱ بهینگی طرح

محققی را که در قسمت قبل توضیح دادیم در نظر بگیرید، او به دنبال کشف مدل صحیح رگرسیون پواسون میان دوز دارو و تعداد کلونی‌های سرطانی بود و برای این هدف بایستی ضرایب مدل را براساس آزمایشاتی برآورد نماید. از آن‌جا که انجام آزمایشات مستلزم صرف هزینه و زمان است؛ او در نظر دارد برای کشف سریع تر مدل صحیح و صرف هزینه کمتر، آزمایشاتی را انجام دهد که برآوردهای به‌دست آمده براساس این آزمایشات بسیار به مقادیر واقعی ضرایب مدل نزدیک باشند، همچنین برای آزمون فرضیاتی در مورد ضرایب مدل بهتر است که برآوردگرها دارای کمترین واریانس باشند و به عبارتی برآوردهای ضرایب رگرسیونی دارای بهینگی در واریانس باشند. بهینگی در واریانس برآوردگرها که آزمایشگر به دنبال آن است، ویژگی طرح‌های بهینه با معیار D- بهینگی است. برای رسیدن به این هدف ابتدا با طرح‌های بهینه و معیارهای بهینگی طرح آشنا می‌شویم.

نخست معرفی خلاصه ای از معیار بهینگی طرح ارائه می‌کنیم. در سال ۱۹۵۹ کیفر و ولفوویتز [۲۳] یک چارچوب نظری برای معیار بهینگی طرح تهیه کردند به این صورت که *طرح را به عنوان اندازه احتمال*، بیان کردند، که به هر نقطه‌ی طرح در فضای طرح، مشاهداتی را تخصیص می‌دهد. دسترسی نظری آن‌ها به بهینگی طرح و معرفی آن‌ها از معیار D و E بهینگی برای مدل رگرسیون خطی منجر به پایه‌ای برای معیارهای بهینگی دیگر مانند معیارهای Q-، G-، F- و A- بهینگی گردید. هر معیار بهینگی طرح، نشان‌دهنده هدف ویژه‌ای است که آزمایش به منظور رسیدن به آن اجرا می‌شود یا یک ویژگی مخصوص که در مدل رگرسیون برازش یافته‌ی نهایی وجود دارد.

ایده‌ی اصلی نظریه‌ی بهینگی طرح این است که استنباط آماری درباره ی کمیت‌های مورد نظر به وسیله **بهینگی** انتخاب سطوح متغیرهای کنترل می‌تواند بهبود پیدا کند. به‌طورکلی، معیار بهینگی طرح می‌تواند به صورت *معیار برآورد یا معیار پیش‌بینی* مشخص شود. یک طرح که از حیث معیار برآورد مانند D- بهینگی، بهینه است، به وسیله مینیمم کردن تغییرات برآوردگرهای پارامتر، اطلاع پارامتر را ماکسیمم می‌کند.

یک طرح که از نظر معیار پیش‌بینی مانند Q- بهینگی، بهینه است، اطلاع درباره ی رویه پاسخ را به وسیله تمرکز روی کمیت پیش‌بینی مدل برازش یافته ماکسیمم می‌کند. در سال ۱۹۸۰، سیلوی [۴۷] و در سال ۱۹۹۲، اتکینسون و دنو[۴] اکثر معیار های بهینگی طرح را که به‌طور معمول استفاده می‌شد، شرح دادند [۵۶]. در این تحقیق، بیشتر روی طرح‌های D- بهینه متمرکز خواهیم شد، لذا D- بهینگی در اینجا با جزییات معرفی می‌شود.

اگر بردار پارامترهای مدل، به وسیله β نشان داده شود، D- بهینگی اشاره می‌کند که انتخاب طرح باید اطلاع در مورد β را به‌وسیله مینیمم کردن واریانس تعمیم‌یافته برآوردگرهای β، ماکسیمم کند. می‌دانیم که روش‌های معمول برای برآورد β در بسیاری از مدل‌های غیر خطی، برآورد درستنمایی ماکسیمم (MLE) است. اگر $\hat{\beta}$ برآورد درستنمایی ماکسیمم β باشد، آنگاه ماتریس واریانس مجانبی $\hat{\beta}$، متناسب با وارون ماتریس اطلاع فیشر است [۲۷]. این ماتریس اطلاع، پایه‌ای را برای معیار بهینگی طرح به وجود می آورد که در (۸.۱) تعریف شده است.

تعریف ۱.۵.۱. اگر شرایط نظم برقرار باشد و بعد β، $p \times ۱$ باشد، ماتریس اطلاع فیشر $I(X,\beta)$ یک ماتریس $p \times p$ است:

$$I(X,\beta) = -E[\frac{\partial^۲ \log(L(X,\beta))}{\partial\beta\partial\acute{\beta}}], \qquad (۸.۱)$$

که در آن $L(X,\beta)$ تابع درستنمایی داده‌ها و Xماتریس طرح است.

توجه کنید که امید ریاضی تحت Y محاسبه می شود.

تعریف ۲.۵.۱. معیار D- بهینگی که واریانس تعمیم‌یافته $\hat{\beta}$ها را مینیمم می‌کند، به‌طور معادل دترمینان ماتریس اطلاع فیشر را ماکسیمم می‌کند. معیار D- بهینگی کلی به صورت زیر تعریف می‌شود:

$$\max_{X\in D} \left| \frac{I(X,\beta)}{n} \right|, \qquad (۹.۱)$$

که در آن D مجموعه ای از تمام طرح‌های ممکن و n اندازه نمونه است. در اکثر مواقع، اندازه نمونه ثابت است، بنابراین، طرح D- بهینه از طریق ماکسیمم کردن دترمینان ماتریس اطلاع فیشر، به‌دست می‌آید.

تعریف ۱.۵.۳. از آنجا که طرح D- بهینه، دترمینان ماتریس اطلاع فیشر را ماکسیمم می‌کند، طبیعی است که D- کارایی طرح دلخواه $\mathbf{X}$ را چنین تعریف کنیم:

$$Eff_D = \left(\frac{\left| I(\mathbf{X}, \beta) \right|}{\left| I(\mathbf{X}^*, \beta) \right|} \right)^{\frac{1}{p}}, \qquad (۱۰.۱)$$

که در آن $\mathbf{X}^*$ طرح D- بهینه با همان اندازه $\mathbf{X}$ است و p تعداد پارامترهای مدل است.

وقتی D- بهینگی به کار برده می‌شود، D- کارایی تعریف شده در (۱۰.۱) عموماً برای مقایسه‌ی طرح‌های متفاوت استفاده می‌شود. گرفتن نسبت دترمینان ها در (۱۰.۱) و به توان $\frac{1}{p}$ رساندن آن‌ها، یک اندازه کارایی را که متناسب با حجم نمونه است و صرف نظر از بعد مدل می‌باشد، نتیجه می‌دهد.

همان‌طور که پیش از این توضیح داده شد، در مدل‌های خطی با واریانس ثابت σ^2، ماتریس اطلاع متناسب با ماتریس واریانس - کوواریانس می‌باشد (بنا به رابطه (۷.۱)) برای کاهش واریانس برآوردگرها در مدل‌های خطی ناچاریم که ماتریس $(\mathbf{X}'\mathbf{X})^{-1}$ و یا به‌طور معادل $\mathbf{I}^{-1}$ را مینیمم نماییم. برای مینیمم کردن وارون ماتریس اطلاع، بایستی دترمینان ماتریس اطلاع را ماکسیمم نماییم، و چون ماتریس اطلاع متناسب با $\mathbf{X}'\mathbf{X}$ است، کافی است دترمینان $\mathbf{X}'\mathbf{X}$ را ماکسیمم نماییم. یعنی x هایی را از فضای طرح پیدا نماییم که دترمینان مورد نظر را ماکسیمم نماید. بنابراین در مدل‌های خطی، ماتریس اطلاع به پارامتر مجهول بستگی ندارد و به آسانی می‌توان دترمینان ماتریس اطلاع را براساس xها ماکسیمم نمود. به عبارت دیگر، ماتریس اطلاع فیشر برای واریانس ثابت مدل خطی به β بستگی ندارد. در نتیجه طرح D- بهینه می‌تواند مستقل از β تعیین و اجرا شود. لکن، برای یک مدل غیر خطی مانند مدل خطی تعمیم‌یافته، ماتریس اطلاع، معمولاً به پارامترهای مدل بستگی دارد. ماتریس اطلاع برای مدل خطی تعمیم‌یافته، به صورت زیر است:

$$I(\mathbf{X}, \beta) = \mathbf{X}'\mathbf{W}\mathbf{X} \qquad (۱۱.۱)$$

که در آن $\mathbf{W}$ ماتریس وزن‌دار هسین[۱] است که به پارامترهای مجهول بستگی دارد

[۱] Hessian

[۲]. برای یک مدل خطی تعمیم‌یافته کانونی مانند مدل رگرسیون لجستیک رابطه (۱۱.۱) به صورت زیر تبدیل می‌شود:

$$I(\mathbf{X}, \beta) = \mathbf{X}'\mathbf{V}\mathbf{X} \qquad (۱۲.۱)$$

که در آن $\mathbf{V}$ یک ماتریس قطری است و (i,i)امین عنصر آن، واریانس iامین متغیر است. اما این واریانس‌ها به β بستگی دارند و لذا ماتریس اطلاع نیز به β وابسته خواهد بود. بنابراین، طرح D- بهینه برای یک مدل خطی تعمیم‌یافته وابسته به پارامتر است. به‌منظور یافتن و اجرای طرح D- بهینه، باید پارامترهای مدل، معلوم باشند. این وابستگی به پارامترهاست که مسأله طرح آزمایش‌ها برای مدل‌های خطی تعمیم‌یافته را پیچیده می‌کند. چندین روش، برای حل این مشکل در کتاب‌ها پیشنهاد شده که در قسمت بعد مروری بر آن‌ها خواهیم داشت.

۶.۱ طرح‌های بهینه برای مدل رگرسیون لجستیک

برای حل مسأله آزمایشگری که روی داروی ضد سرطان کار می‌کند و می‌خواهد برآوردگرهای ضرایب مدل پواسون را با کمترین واریانس به‌دست آورد، نیاز به پیدا کردن طرح D- بهینه برای مدل رگرسیون پواسون داریم. همان‌طور که دیدیم طرح آزمایش‌های بهینه برای مدل‌های خطی تعمیم‌یافته به پارامترهای مجهول بستگی دارد و لذا مسأله ساختاری آن‌ها، لزوماً پیچیده تر از مدل‌های خطی کلاسیک است. در این بخش، برخی از روش‌هایی را که برای غلبه بر این مشکل پیشنهاد شده‌اند، معرفی می‌کنیم. در کتاب‌ها و مقالات، بیشتر تحقیق رایج روی طرح آزمایش‌ها برای مدل‌های خطی تعمیم‌یافته، به مدل رگرسیون لجستیک محدود شده است. بنابراین، در توضیح این روش‌ها، بیشتر در زمینه رگرسیون لجستیک صحبت می‌کنیم.

۱.۶.۱ طرح‌های بهینه موضعی

یک شیوه‌ی ساده‌ی مسأله طرح آزمایش‌ها برای مدل‌های خطی تعمیم‌یافته، قبول کردن بهترین حدس برای پارامترها و سپس پیدا کردن نقاطی است که تابع معیار

بهینگی انتخاب شده طرح را در حدس مورد نظر، ماکسیمم کند. این شیوه منجر به طرحی می‌شود که *طرح بهینه موضعی*[۲] نامیده می‌شود و به وسیله چرنوف[۹] معرفی گردید. بهترین حدس که در طرح بهینه موضعی استفاده می‌شود، ممکن است از آزمایش‌های قبلی بیاید، یا از آزمایشی که مخصوصاً برای این هدف انجام شده، یا صرفاً یک حدس باشد. بدون بحث در نحوه‌ی به‌دست آمدن حدس پارامتر، آن را *برآورد اولیه یا حدس اولیه*، خواهیم نامید.

اکنون، مروری خلاصه بر طرح‌های بهینه موضعی برای مدل رگرسیون لجستیک ارائه می‌کنیم. بیشتر مطالعات بر روی عنوان مدل رگرسیون لجستیک مرتبه اول یک متغیری متمرکز است که به صورت زیر تعریف می‌شود:

$$y \sim Bernoulli(P), \quad P = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x))} \qquad (13.1)$$

در مدل (۱۳.۱)، y پاسخ دودویی (۰ یا ۱) است، که با احتمال p مقدار یک را می‌گیرد و p تابعی مستقل از متغیر x است. برای این مدل، کالیش و رزنبرگر [۲۱] یک طرح بهینه دو سطحی را به‌دست آوردند. آن‌ها تقارن حول نقطه ED_{50} را تعیین کردند (ED مخفف دوز مؤثر است و ED_{100p} بیانگر مقدار xای است که با احتمال p تولید می‌شود و دارای پاسخ یک است). طرح D- بهینه دو سطحی به وسیله قرار دادن نیمی از اجراهای آزمایشی در نقطه $ED_{17/6}$ و نیمی دیگر در نقطه $ED_{82/4}$ به‌دست می‌آید. برای تکمیل طرح D- بهینه باید مقادیر $ED_{17/6}$ و $ED_{82/4}$ معلوم باشد، یا به‌طور معادل، باید اطلاعی در مورد β_0 و β_1 داشته باشیم. در عمل، پارامترها ذاتاً مجهول هستند و نیاز به تهیه‌ی چندین حدس اولیه می‌باشد.

کالیش و رزنبرگر[۲۱]، یک طرح دو سطحی G- بهینه را توسعه دادند که ماکسیمم واریانس احتمال پیش‌بینی $\hat{P}$ را روی ناحیه مورد نظر مینیمم می‌کند. ویلیامز [۵۴] در سال ۱۹۸۶ روی معیار بهینگی دیگری که طول بازه اطمینان را براساس برآوردهای درستنمایی ماکسیم، مینیمم می‌کند، کار کرد. کار دیگر روی طرح‌های بهینه موضعی برای مدل رگرسیون لجستیک شامل کار مایرز و هیز [۱۹] است، آن‌ها طرح‌های D- بهینه و Q- بهینه را برای مدل رگرسیون لجستیک دو متغیری بررسی کردند. اگر

[۲] locally optimal design