سیزدهمین کنفرانس آمار ایران
۴-۲ شهریور ۱۳۹۵

طرح بهینه برای برآورد رگرسیون ستیغی پواسن

صلاح قربانی [۱] ، مهرداد نیاپرست
گروه آمار، دانشگاه رازی کرمانشاه

چکیدهطرح‌های بهینه به عنوان ابزاری که به محقق کمک می‌کند تا پیش‌بینی نتایج را با دقت بیشتری داشته باشد از دیرباز موردتوجه بوده است. اغلب این تحقیقات بر مبنای مدل‌های خطی و با شرایط معمول در مدل‌های خطی از جمله پررتبه بودن ماتریس طرح(رگرسیونی) است. در این مقاله مدل رگرسیونی پواسن به عنوان حالت خاصی از مدل‌های خطی تعمیم یافته به کار برده می‌شود. طرح A-بهینه برای این مدل بر اساس دو برآورد رگرسیون ستیغی پواسن و برآورد رگرسیون پواسن محاسبه می‌گردد. نکته اساسی در استفاده از رگرسیون پواسن ستیغی این است که در این روش می‌توان حالتی را که ماتریس طرح پررتبه نیست در نظر بگیریم. نتایج نشان دادند که استفاده از رگرسیون ستیغی پواسن مناسب‌تر است.

واژه‌های کلیدی: رگرسیون پواسن، رگرسیون ستیغی پواسن و طرح بهینه.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62J07، 62K05، 62J99.

۱ مقدمه

در سال‌های اخیر مسئله‌ی کاهش هزینه و قابل اعتماد بودن نتایج در انجام آزمایش‌های آماری مورد توجه قرار گرفته است. استفاده از طرح‌های بهینه به عنوان راه‌حلی که می‌تواند در راستای هدف گفته شده باشد، توجه آماردانان را به خود جلب کرده است. به طوری که بر اساس نقاط طرح، بهترین برآورد برای پارامترهای مدل به دست آورده شود. بیشتر معیارهایی که برای پیدا کردن طرح بهینه استفاده می‌شود بر اساس برآورد پارامترهای مدل می‌باشد. بنابراین به دست آوردن واریانس برآوردگر پارامترها یا به طور مجانبی عکس ماتریس اطلاع فیشر از اهمیت اساسی برخوردار است.

طرح‌های بهینه[۱] در مدل‌های خطی به طور وسیع مورد توجه آماردانان زیادی قرار گرفته است. وانگ و همکاران (۲۰۰۶) طرح‌های بهینه را برای مدل پواسن به صورت کامل به دست آوردند. در کلیه‌ی مدل‌های فرض شده در وانگ و همکاران (۲۰۰۶) ، فرض بر پررتبه بودن ماتریس طرح(رگرسیونی) بوده است. شرطی که در بعضی موارد ممکن است نقض شود. که این مسئله منجر به ماتریس

[۱] صلاح قربانی: ghorbani.salah@stu.razi.ac.ir

[۱] Optimal Designs

واریانس-کواریانس بدشرطیده[۲] برای برآورد پارامترها می‌شود. در این مقاله، حالتی در مدل پواسن در نظر گرفته می‌شود که شرط پررتبه بودن ماتریس طرح برقرار نباشد. روش رگرسیون ستیغی پواسن[۳] به عنوان راهی برای گریز از این مشکل در مانسون و شوکور (۲۰۱۱) ارائه شده است. ایده‌ی این کار ابتدا توسط هورل و همکاران (۱۹۷۵) ارائه شد. سپس توسط محققان دیگری از جمله لاولس و وانگ (۱۹۷۶) و مونیز و کیبریا (۲۰۰۹) بسط داده شد.

در بخش ۲ به معرفی رگرسیون ستیغی خواهیم پرداخت. در بخش ۳ طرح‌های بهینه شرح داده می‌شود. و بخش ۴ شامل بدست آوردن طرح A-بهینه برای برآورد رگرسیون ستیغی پواسن و مقایسه‌ی آن با برآورد معمولی رگرسیونی پواسن است.

۲ رگرسیون ستیغی پواسن

تعاریف و نمادگذاری این بخش، از مقاله مانسون و شوکور (۲۰۱۱) می‌باشد.

روش رگرسیون ستیغی در ساده ترین شکل مدل، یعنی مدل خطی

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (۱.۲)$$

به صورت زیر می‌باشد که با اضافه کردن عدد کوچک مثبتی ($\mathrm{k} \geq 0$) به مولفه‌های روی قطر اصلی ماتریس $\mathbf{X}^T\mathbf{X}$ از رگرسیون چندگانه بدست آمده است:

$$\hat{\boldsymbol{\beta}}_{\mathbf{RR}} = \left(\mathbf{X}^T\mathbf{X} + \mathrm{k}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

که در آن $\mathbf{X}$ ماتریس طرح و از مرتبه $n \times p$ از مشاهدات است. وقتی که متغیرهای رگرسیونی دارای هم‌خطی[۴] باشند. $\hat{\boldsymbol{\beta}}_{RR}$ بردار پارامترهای برآورد شده رگرسیون ستیغی از مرتبه‌ی $p \times ۱$ است. k عدد مثبتی است. مانسون و شوکور (۲۰۱۱) نشان دادند که میانگین مربعات خطا (MSE)[۵] برای برآورد پارامترها در رگرسیون ستیغی، کوچکتر از مقدار آن برای پارامترهای مدل معمولی با استفاده از روش حداقل مربعات خطا (LSE)[۶] می باشد.

مشخصه‌ی اصلی مدل۱.۲ استفاده از توزیع نرمال برای بیان تغییرات متغیر پاسخ است. موضوعی که ممکن است در مسائل کاربردی با داده‌های شمارشی به راحتی نقض شود. رگرسیون پواسن به عنوان ابزاری که برای چنین داده‌هایی مناسب است می‌تواند به‌کار برده شود. و به صورت زیر

$$y_i \sim P(\mu_i(\beta)), \quad i = ۱, ۲, \ldots, n \qquad (۲.۲)$$

تعریف می شود. که در آن $\mu_i(\boldsymbol{\beta}) = \exp\{x_i^{\mathrm{T}}\boldsymbol{\beta}\}$ یا به عبارت دیگر تابع ربط[۷] آن $log(\mu_i) = x_i^{\mathrm{T}}\boldsymbol{\beta}$ است. x_i بردار پیشگوکننده iم و $\boldsymbol{\beta}_{p\times ۱}$ بردار پارامترهای نامعلوم و ثابت است. لگاریتم تابع درستنمایی و مشتق مرتبه‌ی اول تابع درستنمایی این مدل به شکل زیر

[۲] Ill-Conditioned

[۳] Poisson Ridge Regression

[۴] Multicollinearity

[۵] Mean Squares Error

[۶] Least Squares Error

[۷] Link Function

بدست می‌آید:

$$l(\mu;y) = \sum_{i=1}^{n} y_i \log(\mu_i) - \sum_{i=1}^{n} \mu_i - \log(\prod_{i=1}^{n} y_i!)$$
$$= \sum_{i=1}^{n} y_i \log(\exp\{x_i^T\beta\}) - \sum_{i=1}^{n} \exp\{x_i^T\beta\} - \log(\prod_{i=1}^{n} y_i!).$$

$$\Rightarrow \frac{\partial l}{\partial \beta} = \mathbf{F}^T(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})).$$

که در آن

$$\boldsymbol{\mu}^T(\boldsymbol{\beta}) = (\mu_1(\boldsymbol{\beta}), \mu_2(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta})) \qquad (۳.۲)$$

معادله‌ی ۳.۲ یک معادله‌ی کاملا غیرخطی است. بنابراین برآورد ماکسیمم درستنمایی پارامترهای مدل، با استفاده از روش حداقل مربعات تکراری موزون (IWLS)[۸] عبارت است از:

$$\hat{\boldsymbol{\beta}}_{\mathbf{ML}} = \left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{W}}\hat{\mathbf{Z}}$$

که در آن $\hat{\mathbf{W}} = diag(\hat{\mu}_i(\boldsymbol{\beta}))$ و $\hat{\mathbf{Z}}$ برداری است که مولفه‌های آن به صورت روبرو می‌باشد: $\hat{Z}_i = \log(\hat{\mu}_i(\boldsymbol{\beta})) + \frac{y-\hat{\mu}_i(\boldsymbol{\beta})}{\hat{\mu}_i(\boldsymbol{\beta})}$.
همچنین MSE این برآوردگر برابر است با:

$$E(L_{ML}^2) = E(\hat{\boldsymbol{\beta}}_{\mathbf{ML}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_{\mathbf{ML}} - \boldsymbol{\beta}) = tr[(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\mathbf{X})^{-1}] = \sum_{j=1}^{J} \frac{1}{\lambda_j}$$

که λ_j در آن، iامین مقدار ویژه از ماتریس $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ است.

در مدل رگرسیونی ۲.۲ در صورتی که ماتریس $\mathbf{X}$ پرتبه نباشد، ماتریس $\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)$ بدشرطیده می شود. مانسون و شوکور (۲۰۱۱) از روش رگرسیون ستیغی برای حل این مشکل استفاده نمودند. اگرچه برآورد رگرسیون ستیغی پواسن برخلاف برآورد MLE [۹] رگرسیون پواسن که به طور مجانبی ناریب است، اریب می باشد ولی به راحتی می‌توان نشان داد که نسبت به برآوردگر MLE در این مدل دارای میانگین مربعات خطای (MSE) کمتری است. به پیروی از مانسون و شوکور (۲۰۱۱) در اینجا برآوردگر رگرسیون ستیغی پواسن را با $\hat{\boldsymbol{\beta}}_{PRR}$ نشان می دهیم و عبارت است از:

$$\hat{\boldsymbol{\beta}}_{PRR} = \left(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{X} + \mathrm{K}\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{X}\hat{\boldsymbol{\beta}}_{ML} = \mathbf{Z}\hat{\boldsymbol{\beta}}_{ML}$$

[۸] Iterative Weighted Least Squares

[۹] Minimum Likelihood Estimator

مانسون و شوکور (۲۰۱۱) میانگین مربعات خطای $\hat{\boldsymbol{\beta}}_{PRR}$ را به صورت زیر به دست آوردند:

$$\begin{aligned}
E(L^{2}_{PRR}) &= E\left[(\hat{\boldsymbol{\beta}}_{PRR}-\boldsymbol{\beta})^{T}(\hat{\boldsymbol{\beta}}_{PRR}-\boldsymbol{\beta})\right] \\
&= E\left[\left(\mathbf{Z}\hat{\boldsymbol{\beta}}_{ML}-\mathbf{Z}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\beta}-\boldsymbol{\beta}\right)^{T}\left(\mathbf{Z}\hat{\boldsymbol{\beta}}_{ML}-\mathbf{Z}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\beta}-\boldsymbol{\beta}\right)\right] \\
&= E\left[tr\left[\mathbf{Z}^{T}\mathbf{Z}\left(\hat{\boldsymbol{\beta}}_{ML}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{ML}-\boldsymbol{\beta}\right)^{T}\right]\right]+\boldsymbol{\beta}^{T}(\mathbf{Z}-\mathbf{I})^{T}(\mathbf{Z}-\mathbf{I})\boldsymbol{\beta} \\
&= tr\left[\left(\mathbf{X}^{T}\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{Z}^{T}\mathbf{Z}\right]+k^{2}\boldsymbol{\beta}^{T}\left(\mathbf{X}^{T}\hat{\mathbf{W}}\mathbf{X}+K\mathbf{I}\right)^{-2}\boldsymbol{\beta} \\
&= tr\left(\mathbf{X}^{T}\hat{\mathbf{W}}\mathbf{X}+K\mathbf{I}\right)^{-1}-k.tr\left(\mathbf{X}^{T}\hat{\mathbf{W}}\mathbf{X}+K\mathbf{I}\right)^{-2}+k^{2}\boldsymbol{\beta}^{T}\left(\mathbf{X}^{T}\hat{\mathbf{W}}\mathbf{X}+K\mathbf{I}\right)^{-2}\boldsymbol{\beta} \\
&= \sum_{j=1}^{J}\frac{\lambda_{j}}{(\lambda_{j}+k)^{2}}+k^{2}\boldsymbol{\beta}^{T}\left(\mathbf{X}^{T}\hat{\mathbf{W}}\mathbf{X}+K\mathbf{I}\right)^{-2}\boldsymbol{\beta} \\
&= \tau_{1}(\mathrm{k})+\tau_{2}(\mathrm{k}).
\end{aligned}$$

که در آن $\tau_{1}(\mathrm{k})$ و $\tau_{2}(\mathrm{k})$ به ترتیب برابر با واریانس و توان دوم اریبی می‌باشد.

۳ طرح های بهینه

اگر طرح آزمایش را به صورت مجموعه‌ای از زوج مرتب‌هایی که هر زوج ترکیبی شامل نقطه‌ای از متغیرهای مستقل، و نسبتی که آن نقطه در آزمایش تکرار شده است، در نظر بگیریم، طرح بهینه عبارت است از مجموعه‌ای از این زوج‌ها که منجر به بهترین برآورد(بر اساس معیارهای خاص) برای پارامترهای مجهول می‌شود. معیارهای زیادی برای مقایسه‌ی طرح‌ها و در نهایت پیدا کردن بهترین طرح معرفی شده‌اند. در این قسمت ما از معیاری تحت عنوان A-بهینه استفاده می‌کنیم.

تعریف ۱.۳. طرح A-بهینه: فرض کنید که

$$d=\{(x_{1},p_{1}),...,(x_{k},p_{k})\}$$

با فرض $\sum_{i=1}^{n}p_{i}=1$ و $x_{i}\in\boldsymbol{\chi}$ که $\boldsymbol{\chi}$ فضای متغیر است ، طرحی درون مجموعه‌ی کلیه‌ی طرح‌ها باشد. در این صورت

$$d^{*}=\{({x_{1}}^{*},p_{1}),...,({x_{k}}^{*},p_{k})\}$$

را طرح A-بهینه گویند اگر

$$d^{*}=arg\min_{d}tr(\mathrm{var}(\hat{\boldsymbol{\beta}}_{d}))$$

از آنجایی که $\hat{\boldsymbol{\beta}}$ وابسته به طرح d است بنابراین به جای $\hat{\boldsymbol{\beta}}$ از $\hat{\boldsymbol{\beta}}_{d}$ استفاده شده است.

در مقایسه‌ی طرح‌های بهینه برای برآورد رگرسیون ستیغی پواسن، با برآورد رگرسیون معمولی پواسن، از معیار کارایی[۱۰] به صورت

$$\mathbf{A_{eff}}=\frac{\mathrm{tr}(\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{d^{*},PRR}}))}{\mathrm{tr}(\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{d^{*},PR}}))}$$

استفاده می‌شود. به طوری که $\hat{\boldsymbol{\beta}}_{d^{*},PRR}$ و $\hat{\boldsymbol{\beta}}_{d^{*},PR}$ به ترتیب برآوردگر رگرسیون ستیغی پواسن و برآوردگر رگرسیون معمولی پواسن بر اساس طرح A-بهینه است.

[۱۰] Efficiency Criteria

۴ مثال عددی

مدل رگرسیون پواسن با میانگین $\mu_i(\boldsymbol{\beta}) = e^{\beta_0 + \beta_1 x_i}$ را در نظر بگیرید، که حالت خاصی از مدل ۲.۲ است. همچنین فرض شود که مجموعه‌ی طرح‌های در نظر گرفته برای این مدل، طرح‌های اشباع شده باشد. به عبارت دیگر طرح‌هایی با دو نقطه در تکیه‌گاه طرح را در نظر می‌گیریم:

$$d = \{(x_1, p), (x_2, 1-p)\}$$

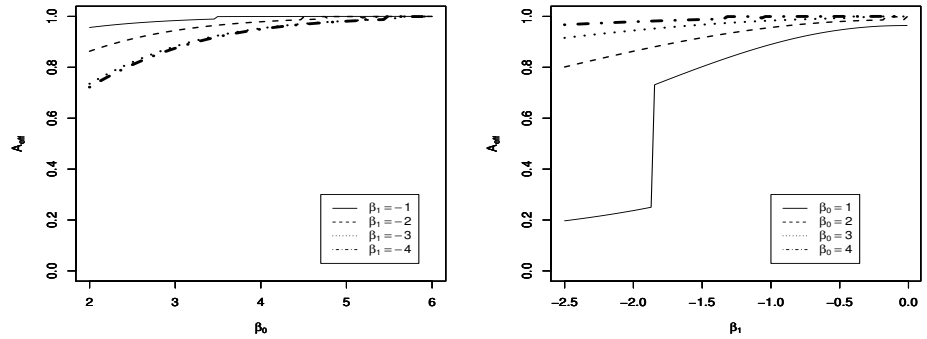
نتایج طرح‌های A-بهینه برای بعضی مقادیر بردار $\boldsymbol{\beta}$ در جدول ۱ آمده است. نتایج نشان می‌دهد که در تمام نقاط انتهایی بردار $\boldsymbol{\beta}$ ، یک نقطه از طرح، ثابت و به ازای گروه کنترل[11] می‌باشد. همچنین با افزایش مقدار β_1 و به ازای مقادیر ثابت β_0 ، نسبت مشاهدات در گروه کنترل افزایش می‌یابد. البته نقاط طرح بهینه در دو روش، به هم نزدیک بوده و این امر به این دلیل رخ داده که در اینجا امکان اینکه ماتریس طرح پررتبه باشد، کم است.

جدول ۱: مقایسه‌ی روش رگرسیون معمولی پواسن و رگرسیون ستیغی پواسن

مقادیر اولیه‌ی پارامترهای مدل		رگرسیون ستیغی پواسن			رگرسیون پواسن		
β_0	β_1	x_1	x_2	p	x_1	x_2	p
۱	-۰.۵	۰	۴.۲۵	۰.۵۹	۰	۴.۳۱	۰.۶
۱	-۱	۰	۲.۱۷	۰.۴۶	۰	۲.۲۶	۰.۴۴
۱	-۱.۵	۰	۱.۴۴	۰.۴۲	۰	۱.۵۵	۰.۳۶
۱	-۲	۰	۱.۰۳	۰.۴۱	۰	۱.۱۹	۰.۳۲
۲	-۰.۵	۰	۴.۲۹	۰.۵۸	۰	۴.۳۱	۰.۶
۲	-۱	۰	۲.۲۳	۰.۴۵	۰	۲.۲۶	۰.۴۴
۲	-۱.۵	۰	۱.۵۲	۰.۳۹	۰	۱.۵۶	۰.۳۷
۲	-۲.۵	۰	۰.۹۲	۰.۳۴	۰	۰.۹۷	۰.۳

مقایسه‌ی طرح‌های A-بهینه‌ی مختلف بر اساس دو برآوردگر رگرسیونی پواسن و برآوردگر رگرسیون ستیغی پواسن در شکل ۱ آمده است که نشان از بهتر بودن طرح A-بهینه برای برآورد رگرسیون ستیغی پواسن است.

[11] Control Group



شکل ۱: A -کارایی طرح های بهینه برای مدل رگرسیون ستیغی پواسن در مقایسه با مدل رگرسیون پواسن معمولی. سمت چپ: A -کارایی برحسب β_0 برای مقادیر مختلف β_1 . سمت راست: A -کارایی برحسب β_1 برای مقادیر مختلف β_0

بحث و نتیجه‌گیری

نتایج نشان می‌دهند که در حالتی‌که از رگرسیون پواسن ستیغی استفاده شود، مقدار **مجموع** واریانس برای پارامترها کمتر از **مجموع** واریانس پارامترهای مدل رگرسیون پواسن است. همچنین با توجه به محاسبه‌ی A-کارایی می‌توان نتیجه گرفت که کارایی استفاده از برآورد رگرسیون ستیغی پواسن، بالاتر از رگرسیون پواسن است.

مراجع

Hoerl, A.E., Kennard, R.W., Baldwin, K.F., (1975). Ridge regression: some simulation. *Communications in Statistics—Theory and Methods*. **4**, 105–123.

Lawless, J.F., Wang, P., (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics—Theory and Methods*. **5** , 307–323.

Månsson, K., & Shukur, G. (2011). A Poisson ridge regression estimator. *Economic Modelling*, **28(4)**, 1475-1481.

Muniz, G., Kibria, B.M.G., (2009). On some ridge regression estimators: An Empirical Comparisons. *Communications in Statistics—Simulation and Computation*. **38**, 621–630.

Wang, Y., Myers, R. H., Smith, E. P., & Ye, K. (2006). D-optimal designs for Poisson regression models. *Journal of statistical planning and inference*,**136(8)**, 2831-2845.