مطالعه رفتار حدی برآوردگرهای انقباضی در مدل رگرسیون تاوانیده با نرم مستطیلی

مینا نوروزی‌راد، محمد آرشی

گروه آمار، دانشگاه صنعتی شاهرود

چکیده: برآوردگرهای تاوانیده در سال‌های اخیر در برآورد پارامترهای رگرسیونی بسیار مورد توجه قرار گرفته‌اند، که معروف‌ترین آن‌ها برآوردگرهای تاوانیده با نُرم مستطیلی هستند. این برآوردگرها، همزمان انتخاب متغیر و برآورد پارامتر انجام می‌دهند. در این مقاله، با استفاده از اطلاعات پیشین غیرقطعی در مورد پارامترها، برآوردگرهای بهتری با مخاطره کمتر در مقایسه با برآوردگر لاسو، تاوانیده با نُرم مستطیلی ارائه شده است. برتری کارآیی برآوردگرهای انقباضی پیشنهاد شده در یک مطالعه شبیه‌سازی نسبت به برآوردگر لاسو نشان داده شده است. همچنین کاهش در مقادیر میانگین خطاهای پیش‌بینی در مجموعه داده‌های سرطان آمار و ارقام ایالات متحده آمریکا حاکی از قدرت پیش‌گویی برآوردگرهای انقباضی است.

واژه‌های کلیدی: برآوردگر انقباضی، برآوردگر بهبودیافته، برآوردگر لاسو، توزیع مجانبی، خطای پیش‌گویی، نُرم مستطیلی.

°آدرس الکترونیک مسئول مقاله: مینا نوروزی‌راد، mina.norouzirad@gmail.com
°کد موضوع‌بندی ریاضی (۲۰۱۰): 62F12، 62J07، 62F30

۱ مقدمه

مدل‌های آماری به‌کار گرفته می‌شوند تا بینشی درباره کمیت‌های مجهول (پارامترها) به‌دست آید. با این وجود، در بسیاری از موقعیت‌ها، آماردان‌ها علاوه بر اطلاعات به‌دست آمده از نمونه، از سایر اطلاعات نیز در برآورد پارامترها استفاده می‌کنند. منظور از سایر اطلاعات در اینجا می‌تواند اطلاعات غیرنمونه‌ای[۱] (NS) یا اطلاعات پیشین غیرقطعی[۲] (UP) درباره پارامترهای مورد علاقه باشد. به‌کار بردن این اطلاعات غیرنمونه‌ای در برآورد پارامتر، به‌ویژه زمانی‌که اطلاعات بر اساس نمونه ممکن است محدود باشد، مفید است.

استفاده از اطلاعات قابل اعتماد معمولا نتایج پرباری به همراه خواهد داشت. اما در بعضی موارد تجربی، از صحت اطلاعات NS یا UP اطمینان نداریم. به‌عنوان مثال، داده‌هایی را در نظر بگیرید که از اندازه‌گیری نوعی تومور در موش‌ها در زمان‌های مختلف بعد از تزریق مواد سرطان‌زا به‌دست آمده‌اند. معمولا فرض می‌شوند این داده‌ها از آزمایش‌ها و مطالعات بالینی به‌دست آمده است. زیست‌شناسان علاقه‌مندند پارامتر نرخ رشد λ را در حالتی که بر اساس آزمایشات قبلی به مقدار λ_0، که یک اطلاع UP است، دست یافته‌اند، برآورد کنند. اما با توجه به اینکه در شرایط محیطی مختلف رفتار سلول‌ها ممکن است متفاوت باشد، مقدار λ_0 می‌تواند تغییر کند؛ بنابراین زیست‌شناسان با وجود اینکه برای در نظر گرفتن λ_0 به عنوان مقدار واقعی پارامتر نرخ رشد در آزمایشات بالینی دلایل کافی دارند، نسبت به این مقدار از اطمینان مناسبی برخوردار نیستند.

در واقع، اغلب در علوم کاربردی، یک آزمایش با داشتن پیش‌زمینه‌ای از نتیجه آن اجرا می‌شود. استفاده از قوانین استاندارد آماری، که معمولا فرض می‌شود هیچ شناختی از نتیجه آزمایش ندارند، ممکن است مفید نباشند.

در حالت کلی، نتایج و عواقب استفاده از اطلاعات غیرنمونه‌ای وابسته به کیفیت یا اعتبار اطلاعاتی است که در برآورد از آن استفاده می‌شود. این اطلاع پیشین غیرقطعی به شکل فرضیه صفر می‌تواند به دو روش در برآورد کردن مورد استفاده قرار گیرد. در روش اول، یک آزمون اولیه بر روی اعتبار اطلاع پیشین غیرقطعی به شکل یک محدودیت بر روی فضای پارامتر انجام می‌شود و بر اساس نتیجه این آزمون می‌توان بین برآوردگر بر اساس اطلاع UP و برآوردگر بدون استفاده از اطلاع UP، یکی را انتخاب کرد. این ایده، اولین بار توسط بنکرافت (۱۹۴۴) مطرح شد.

روش دیگر، به‌کار بردن ایده جیمز و استاین است. استاین (۱۹۵۶) و جیمز و استاین (۱۹۶۱) نشان

[۱] Nonsample
[۲] Uncertain prior (UP)

دادند که میانگین نمونه در مدل توزیع نرمال p-متغیره برای تابع زیان درجه دوم به ازای $p \geq ۳$ برآوردگر پذیرفتنی[۳] نیست و این نتیجه منجر به ایجاد کلاس بزرگی از برآوردگرهای انقباضی شد که به افتخار پروفسور چارلز استاین، برآوردگرهای انقباضی نوع استاین نامیده شدند. در دهه‌های اخیر، محققین توجه قابل ملاحظه‌ای به برآوردگرهای نوع استاین در نمونه‌های کوچک و بزرگ داشته‌اند. جاج و باک (۱۹۷۸) برآوردگرهای آزمون اولیه و نوع استاین را مورد مقایسه قرار دادند. صالح (۲۰۰۶) برآوردگرهای نوع استاین را دوباره‌نویسی کرد و برآوردگرهای مجانبی و ناپارامتری آن را نیز ارائه نمود. با توجه به تاثیر بسیار زیاد نتیجه استاین (۱۹۵۶)، تعداد مقالاتی که کاربرد این برآوردگرها را در علوم مختلف نشان می‌دهد، افزایش یافت. برای جزئیات بیشتر درباره برآوردگرهای نوع استاین، منابع صالح (۲۰۰۶)، شالاب (۱۹۹۸)،شالاب (۲۰۰۱)، گروبر (۱۹۸۸)، صالح و کیبریا (۲۰۱۱)، کیبریا و صالح (۲۰۰۳)، کیبریا و صالح (۲۰۱۲)، رحیم و احمد (۲۰۱۱) و همچنین احمد و نیکل (۲۰۱۲)، لیان (۲۰۱۳)، صالح و همکاران (۲۰۱۴) و تولومیس (۲۰۱۵) را ببینید.

مدل رگرسیون خطی

$$\boldsymbol{y}_n = X_n\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (۱)$$

را در نظر بگیرید که در آن $\boldsymbol{y}_n$ بردار تصادفی $n \times ۱$ از متغیرهای پاسخ، X_n ماتریس طرح غیر تصادفی $n \times p$ ($n > p$)، $\boldsymbol{\beta} = (\beta_۱, \dots, \beta_p)^T$ بردار پارامترهای رگرسیونی و $\boldsymbol{\epsilon} = (\epsilon_۱, \dots, \epsilon_n)^T$ بردار آشفتگی شامل متغیرهای تصادفی مستقل و هم‌توزیع با تابع توزیع F هستند. همچنین، $E(\boldsymbol{\epsilon}) = ۰$ و $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma^۲ I_n$، که در آن $\sigma^۲$ پارامتر مجهول مثبت، متناهی و I_n ماتریس همانی $n \times n$ است.

معمولا بردار پارامتر رگرسیونی $\boldsymbol{\beta}$ به‌گونه‌ای برآورد می‌شود که تابع زیان، تابعی از اختلاف بین مقدار واقعی $\boldsymbol{y}$ و مقدار پیش‌بینی‌شده $\hat{\boldsymbol{y}}$، کمینه شود. به‌طور معمول از تابع زیان توان دوم خطا به‌صورت

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^۲ = \sum_{i=۱}^{n}(y_i - \hat{y}_i)^۲$$

استفاده می‌شود. برآوردگری که به این روش به‌دست می‌آید، برآوردگر کمترین توان‌های دوم نام دارد. گاهی ممکن است در واقعیت یکی از مولفه‌های $\boldsymbol{\beta}$ مقدار بزرگی داشته باشد. بزرگی این مقدار، افزایش تابع زیان را به دنبال خواهد داشت. چون احتمال این‌که مولفه متناظر برآوردگر به اندازه مقدار پارامتر، بزرگ باشد، تقریبا صفر است، لذا کمینه کردن تابع زیان منجر به بیش برازشی می‌شود. یکی از روش‌هایی که می‌توان در برطرف کردن این مشکل به‌کار گرفت این است که به اصطلاح پارامتر تاوانیده شود. تاوان‌های متفاوتی بر حسب نیاز

[۳] Admissible

می‌تواند در نظر گرفته شود که در هر حالت، برآوردگر نام خاصی به خود می‌گیرد. تاوان‌های $\sum_{j=1}^{p}|\beta_j|$ و $\sum_{j=1}^{p}\beta_j^2$ از معروف‌ترین اعضای این گروه از توابع هستند که به‌ترتیب بر اساس نُرم‌های مستطیلی (L_1) و اقلیدسی (L_2) به‌دست آمده‌اند. لازم به ذکر است که به نُرم L_1، منهتن نیز می‌گویند. در ادامه، به تابع‌های تاوان بر پایه‌ی نُرم L_1 و L_2 به ترتیب تاوان‌های L_1 و L_2 گفته می‌شود.

مشهورترین عضو از کلاس برآوردگرهای تاوانیده، برآوردگر ریج است. هورل و کنارد (۱۹۷۰) برآوردگر رگرسیونی «ریج» را معرفی کردند که دروازه ورود به دنیای «برآوردگرهای تاوانیده» بر اساس روش منظم‌سازی تیکونوف (۱۹۶۳) بود. رگرسیون ریج، مقدمه‌ای برای راهیابی به دنیای مسئله برآورد و انتخاب متغیر است. این رگرسیون، با مشکل هم‌خطی در مدل‌های خطی مبارزه می‌کند و بر این اساس برآورد تاوانیده متولد شد. با توجه به حضور همه متغیرها در برآوردگر ریج، تفسیر پذیری آن به سادگی امکان‌پذیر نیست. عضو دیگری از این کلاس، برآوردگر لاسو (تیبشیرانی، ۱۹۹۶) است. تیبشیرانی (۱۹۹۶) به‌جای استفاده از تاوان L_2 توان‌های دوم خطا را نسبت به تابع تاوان L_1 کمینه نمود. این برآوردگر منجر به پیدایش برآوردگرهای جدیدی مانند برآوردگر مشتق‌پذیر کوتاه‌شده هموار[۴] (فن و لی، ۲۰۰۱)، الاستیک‌نت (ژو و هیستی، ۲۰۰۵)، لاسوی سازوار (ژو، ۲۰۰۶) و لاسوی آستانه سخت (بلونی و چرنوکوف، ۲۰۱۳) گردید. استفاده از تاوان L_1، هر ضریب را به سمت صفر منقبض کرده و متغیرهای اضافی را دقیقا صفر می‌کند. در واقع، برآوردگر لاسو، همزمان هم انتخاب متغیر انجام می‌دهد و هم ضرایب را منقبض می‌کند. یک کاربرد جالب از برآوردگرهای لاسو در مدل‌های تنک[۵] (مدل‌هایی که تعداد پارامترهای صفر آن زیاد باشد) است. کاربرد دیگر برآوردگر لاسو، زمانی است که بعد فضای پارامتر بیشتر از بعد فضای نمونه باشد.

در ساختار برآوردگر لاسو، یک پارامتر اساسی به نام حد آستانه وجود دارد که درجه همواری را کنترل می‌کند. انتخاب پارامترهای هموارسازی بخش مهمی از برازش مدل است و در کاربردهای آماری نقش مهمی ایفا می‌کند. عموما از دو روش برای انتخاب این پارامتر استفاده می‌شود: (الف) اعتبارسنجی متقابل[۶] و (ب) معیارهای اطلاع مانند AIC و BIC که اغلب مدل‌های کارا را پیشنهاد می‌دهند.

هنسن (۲۰۱۶) مقایسه‌ای بین برآوردگرهای لاسو، انقباضی نوع استاین و بهترین انتخاب بر اساس مخاطره کمترین توان‌های دوم انجام داد. وی نتیجه گرفت که برآوردگر لاسو و برآوردگر نوع استاین به‌طور یکنواخت بر دیگری برتری ندارد. هم‌چنین در یک مطالعه شبیه‌سازی نشان داد که برآوردگر لاسو به پارامترهای مدل نسبتا

[۴] Smoothly Clipped Absolute Derivation (SCAD)
[۵] Sparse models
[۶] Cross validation

حساس است و در بخشی از فضای پارامتر، برآوردگر لاسو مخاطره نسبتا بیشتری نسبت به برآوردگر کمترین توان‌های دوم دارد.

حال، شرایطی را در نظر بگیرید که هدف، کمینه کردن توان‌های دوم خطا نسبت به تابع تاوان L_1 با در دسترس داشتن اطلاع UP باشد. برای استفاده از روش اول لازم است بین دو برآوردگر با اطلاع UP و بدون اطلاع UP، انتخاب مناسبی صورت گیرد. در این مقاله، در راستای یافتن راهکار مناسبی برای این مسئله، برآوردگرهای لاسو با اطلاع پیشین غیرقطعی و لاسوی نوع استاین معرفی می‌شود. سپس برآوردگرهای جدید با برآوردگر لاسوی تیبشیرانی از دو منظر کارآیی مجانبی و خطای پیش‌گویی مقایسه می‌شوند.

ساختار مقاله حاضر در بخش‌های بعدی به‌صورت زیر است. در بخش ۲، مدل رگرسیونی خطی و استراتژی‌های مختلف برآورد پارامتر $\boldsymbol{\beta}$ مورد بررسی قرار می‌گیرد. توزیع‌مجانبی برآوردگرها در بخش ۳ ارائه شده است. همچنین در بخش ۴، با استفاده از شبیه‌سازی و مثال عددی به مقایسه برآوردگرها پرداخته شده است. بخش ۵ نیز به بحث و نتیجه‌گیری اختصاص دارد.

۲ برآوردگرها

فرض کنید اطلاعات غیر نمونه‌ای درباره پارامتر $\boldsymbol{\beta}$ را بتوان به‌صورت

$$H\boldsymbol{\beta} = \boldsymbol{h} \qquad (۲)$$

بیان کرد که H یک ماتریس $q \times p$ با رتبه q $(q \leq p)$ و $\boldsymbol{h}$ یک بردار $q \times 1$ از ثابت‌های معلوم است. q تعداد محدودیت‌های خطی است که باید مورد آزمون قرار بگیرد. رتبه H، q است، یعنی محدودیت‌ها مستقل خطی هستند. این محدودیت‌ها ممکن است (الف) یک واقعیت ناشی از ملاحظات تئوری یا تجربی باشد، (ب) یک فرضیه که باید مورد آزمون قرار بگیرد و (ج) یک شرط اضافی مصنوعی که فزونگی را در توصیف یک مدل کاهش یا حذف می‌کند، باشند. (سن‌گوپتا و جامالاداکا، ۲۰۰۳).

در حالت کلی، هدف اصلی در مدل رگرسیون چندگانه (۱) برآورد پارامترها و پیش‌گویی متغیر پاسخ برای یک ماتریس طرح دلخواه است. در این بخش روش‌های مختلف برآورد پارامتر $\boldsymbol{\beta}$ پیشنهاد شده است.

۱- برآوردگر کمترین توان‌های دوم (OLS)

معمول‌ترین روش برآورد بردار پارامتر $\boldsymbol{\beta}$ در (۱) روش کمترین توان‌های دوم (OLS) است. در این روش پارامترها به گونه‌ای برآورد می‌شوند که مجموع توان‌های دوم مانده‌ها،

$$\|\boldsymbol{y}_n - X_n\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

کمینه شود. به شرطی که ماتریس طرح X_n رتبه کامل ستونی باشد و $X_n^TX_n$ معکوس‌پذیر باشد، برآوردگر کمترین توان‌های دوم پارامتر $\boldsymbol{\beta}$ به‌صورت

$$\tilde{\boldsymbol{\beta}}_n = C_n^{-1}X_n^T\boldsymbol{y}_n, \qquad C_n = X_n^TX_n$$

به‌دست می‌آید. به‌طور متناظر، برآوردگر σ^2 به‌صورت

$$s^2 = \frac{1}{m}(\boldsymbol{y}_n - X_n\tilde{\boldsymbol{\beta}}_n)^T(\boldsymbol{y}_n - X_n\tilde{\boldsymbol{\beta}}_n); \quad m = n - p$$

حاصل می‌شود. واضح است که $\tilde{\boldsymbol{\beta}}_n \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2C_n^{-1})$ از متغیر ms^2/σ^2 با توزیع کی‌دو مرکزی با m درجه آزادی مستقل است.

۲- برآوردگر لاسو

تیبشیرانی (۱۹۹۶) روشی جدید برای انتخاب متغیرها پیشنهاد کرد که در این روش، مدل‌های دقیق، پایا و صرفه‌جو[۷] (تفسیرپذیرتر) را معرفی کرد و نام آن را لاسو (عملگر انتخاب و کمترین قدر مطلق انقباضی[۸]) نامید. برآوردگر لاسو، نسخه تاوانیده برآوردگر OLS است. بر اساس ویژگی تنکی نرم L_1، برآوردگر لاسو در سال‌های اخیر مورد توجه قرار گرفته است. برآوردگر لاسوی تیبشیرانی (۱۹۹۶) از حل مسئله بهینه‌سازی

$$\hat{\boldsymbol{\beta}}_n^L = \min_{\boldsymbol{\beta}}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}X_{ij}\beta_j)^2\}, \quad \sum_{i=1}^{p}|\beta_j| \leq t$$

به‌دست می‌آید، که در آن t یک مقدار ثابت است. اگر $t = 0$، مدل فقط شامل عرض از مبدا است در حالی‌که اگر $t = \infty$، آنگاه مدل کامل خواهد بود. اگر $t > \sum_{i=1}^{p}|\beta_i^*|$ که $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ یک برآوردگر اولیه برای $\boldsymbol{\beta}$ است و معمولا این برآوردگر را $\tilde{\boldsymbol{\beta}}_n$ در نظر می‌گیرند، روش لاسو منجر به برآوردگر OLS می‌شود. با این وجود، اگر $0 < t < \sum_{j=1}^{p}|\tilde{\beta}_j|$، آنگاه مسئله بهینه‌سازی فوق معادل با

$$\hat{\boldsymbol{\beta}}_n^L = \text{argmin}_{\boldsymbol{\beta}}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}X_{ij}\beta_j)^2 + \lambda_n\sum_{j=1}^{p}|\beta_j|\}, \quad \lambda_n \geq 0$$

[۷] Parsimonious
[۸] Least absolute shrinkage and selection operator (LASSO)

است، که در آن λ_n پارامتر حد آستانه است که سطح تنکی (تعداد پارامترهای صفر) را در برآوردگر لاسو کنترل می‌کند. از آن جایی که این برآوردگر وابسته به اطلاعات پیشین غیرقطعی نیست، از این پس این برآوردگر را برآوردگر محدود نشده لاسو[۹] (ULE) نامیده و از نماد $\hat{\boldsymbol{\beta}}_n^{UL}$ برای نمایش آن استفاده می‌شود. لازم به ذکر است در ادامه برآوردگرهای جدیدی بر اساس این برآوردگر ارائه می‌شوند.

۳- برآوردگر محدود شده

تا این‌جا، فرض شده است اطلاعات تنها وابسته به نمونه است و اطلاعات غیرنمونه‌ای هیچ تاثیری در برآورد ندارد. با این وجود، در برخی موارد ممکن است که اطلاعات غیرنمونه‌ای (یک محدودیت پیشین روی فضای پارامتر) روی پارامترهای مدل به عنوان محدودیت وجود داشته باشد.

چنان‌چه برآوردگر لاسو در رابطه (۲) صدق کند، برآوردگر لاسوی محدودشده[۱۰] (RLE) نامیده شده و با $\hat{\boldsymbol{\beta}}_n^{RL}$ نشان داده می‌شود. مشابه برآوردگر محدودشده OLS پارامتر $\boldsymbol{\beta}$، نسبت به محدودیت $H\boldsymbol{\beta} = \boldsymbol{h}$ (رابطه ۴-۱-۷ در صالح (۲۰۰۶) را ببینید) برآوردگر لاسوی محدودشده به‌صورت

$$\hat{\boldsymbol{\beta}}_n^{RL} = \hat{\boldsymbol{\beta}}_n^{UL} - C_n^{-1}H^T(HC_n^{-1}H^T)^{-1}(H\hat{\boldsymbol{\beta}}_n^{UL} - \boldsymbol{h}) \qquad (۳)$$

پیشنهاد می‌شود. اگر محدودیت فرض شده در رابطه (۲) برقرار باشد، $\hat{\boldsymbol{\beta}}_n^{RL}$ بر برآوردگر $\hat{\boldsymbol{\beta}}_n^{UL}$ برتری دارد. به عبارت دقیق‌تر، مخاطره نسبت به زیان L_2 آن کمتر است.

لازم به ذکر است که اگر ادعای مطرح شده در مورد پارامترها (اطلاعات غیرنمونه‌ای) با قطعیت کامل صحیح باشد، دیگر لزومی به استفاده از روش‌های آماری وجود نخواهد داشت. اما در واقعیت، همیشه عدم قطعیت وجود دارد و در این راستا، نمی‌توان به تعریف برآوردگر محدودشده به عنوان تنها برآوردگری که برآوردگر محدود نشده را بهبود می‌بخشد، اکتفا کرد. هدف، یافتن برآوردگری است که برآوردگر محدود نشده را با در نظر گرفتن عدم قطعیت ادعای مطرح شده، بهبود بخشد. بنابراین، برآوردگر محدودنشده به منظور تعریف سایر برآوردگرها معرفی می‌شود و یک برآوردگر کاربردی نخواهد بود.

[۹] Unrestricted LASSO estimator (ULE)
[۱۰] Restricted LASSO Estimator (RLE)

۴- برآوردگر آزمون اولیه

اطلاعات پیشین به شکل (۲) ممکن است با تغییر دادن فضای پارامتر وارد مدل شود. در این مورد، فضای پارامتر (محدود شده) جدید، یک زیر فضا از فضای کلی است. در حالت کلی، کاهش در بعد بردار پارامتر، برآورد کارآیی را ارائه می‌دهد. با این وجود، اگر محدودیت برقرار نباشد، چنین نتیجه‌گیری طبیعتا به‌دست نمی‌آید. به این دلیل، استفاده از روش فیشر و ارائه یک برآوردگر آزمون اولیه لاسو[۱۱] (PTLE) با انتخاب $\hat{\boldsymbol{\beta}}_n^{UL}$ یا $\hat{\boldsymbol{\beta}}_n^{RL}$ بر اساس قبول یا رد فرضیه صفر $\mathcal{H}_o : H\boldsymbol{\beta} = \boldsymbol{h}$ راهکاری مناسب می‌باشد. این برآوردگر به‌صورت

$$\hat{\boldsymbol{\beta}}_n^{PL} = \hat{\boldsymbol{\beta}}_n^{UL} - (\hat{\boldsymbol{\beta}}_n^{UL} - \hat{\boldsymbol{\beta}}_n^{RL}) I(\mathcal{L}_n \leq \mathcal{L}_{n,\alpha}) \qquad (۴)$$

خواهد بود، که در آن $\mathcal{L}_{n,\alpha}$ چندک αام توزیع دقیق آماره‌ی آزمون $\mathcal{L}_n$، تحت فرضیه $\mathcal{H}_o : H\boldsymbol{\beta} = \boldsymbol{h}$ است.

دو پیشنهاد برای آماره آزمون وجود دارد. بر اساس روش پیشنهادی صالح (۲۰۰۶) آماره آزمون می‌تواند به‌صورت

$$\mathcal{L}_n = \frac{(H\tilde{\boldsymbol{\beta}}_n - \boldsymbol{h})^T (HC_n^{-1}H^T)^{-1} (H\tilde{\boldsymbol{\beta}}_n - \boldsymbol{h})}{s^2} \qquad (۵)$$

باشد که در این مقاله از این آماره آزمون که بر پایه برآوردگر OLS می‌باشد، استفاده شده است.

لازم به ذکر است این آزمون می‌تواند بر اساس برآوردگر لاسو نیز نوشته شود. باید در نظر داشت با وجودی‌که استفاده از برآوردگر لاسو در آماره آزمون باعث می‌شود محاسبات ساده‌تر شود اما هیچ‌گونه راهکاری برای یافتن توزیع مجانبی برآوردگرهای پیشنهادی بر اساس این آماره آزمون در دسترس نیست.

۵- برآوردگر انقباضی

کارآیی برآوردگرهای محدودشده و آزمون اولیه شدیدا به کیفیت اطلاعات UP (برقراری یا میزان عدم برقراری محدودیت) بستگی دارد. از طرف دیگر، هر چه از فرضیه صفر فاصله گرفته شود، برآوردگرهای انقباضی در مقایسه با سایر برآوردگرها کمترین مخاطره را دارند. همان‌طور که مشاهده می‌شود برآوردگر آزمون اولیه شدیدا وابسته به سطح معنی‌داری α است و رفتار گسسته‌ای دارد که نشان می‌دهد یکی از حالات فرین برآوردگر محدود نشده لاسو یا برآوردگر محدودشده لاسو را بر اساس نتیجه آزمون می‌پذیرد. بر این اساس، استفاده از یک برآوردگر پیوسته و مستقل از α ممکن است باعث افزایش کارآیی شود. برای این منظور ایده جیمز و استاین

[۱۱] Preliminart Test LASSO Estimator (PTLE)

(۱۹۶۱) و لاسو را ترکیب کرده و برآوردگر انقباضی نوع استاین لاسو[۱۲] (SSLE) به صورت

$$\hat{\boldsymbol{\beta}}_n^{SL} = \hat{\boldsymbol{\beta}}_n^{UL} - k_n(\hat{\boldsymbol{\beta}}_n^{UL} - \hat{\boldsymbol{\beta}}_n^{RL})\mathcal{L}_n^{-1}, \quad k_n = \frac{m(q-2)}{(m+2)}, m = n - p \qquad (۶)$$

پیشنهاد می‌شود که در آن k_n ثابت انقباضی است و در یک بازه به گونه‌ای انتخاب می‌شود که $\hat{\boldsymbol{\beta}}_n^{SL}$ بر برآوردگر $\hat{\boldsymbol{\beta}}_n^{UL}$ برتری داشته باشد.

۶- برآوردگر آزمون اولیه بهبودیافته

وقتی فرضیه صفر تقریبا درست باشد برآوردگرهای محدودشده و آزمون اولیه بهتر از برآوردگرهای انقباضی هستند، حال سوال اینجاست که در این حالت آیا می‌توان برآوردگر دیگری تعریف کرد که برآوردگرهای آزمون اولیه و محدود شده را بهبود بخشد. با توجه به اینکه در زیرفضایی از پارامتر، برآوردگر انقباضی بهتر از برآوردگر محدودنشده عمل می‌کند، اگر در برآوردگر آزمون اولیه، $\hat{\boldsymbol{\beta}}_n^{SL}$ با $\hat{\boldsymbol{\beta}}_n^{UL}$ جایگذاری شود، برآوردگر به‌دست آمده مخاطره کمتری نسبت به برآوردگر آزمون اولیه خواهد داشت. سالو و همکاران (۱۹۷۲) بر این اساس برآوردگر آزمون اولیه بهبودیافته را بر پایه برآوردگر OLS پیشنهاد دادند. برآوردگر آزمون اولیه بهبودیافته لاسو[۱۳] (IPTLE) به‌صورت

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n^{IPL} &= \hat{\boldsymbol{\beta}}_n^{SL} - (\hat{\boldsymbol{\beta}}_n^{SL} - \hat{\boldsymbol{\beta}}_n^{RL}) I(\mathcal{L}_n \leq \mathcal{L}_{n,\alpha}) \\ &= \hat{\boldsymbol{\beta}}_n^{PL} - k_n \mathcal{L}_n^{-1} I(\mathcal{L}_n > \mathcal{L}_{n,\alpha})(\hat{\boldsymbol{\beta}}_n^{UL} - \hat{\boldsymbol{\beta}}_n^{RL}) \end{aligned} \qquad (۷)$$

پیشنهاد می‌شود. با نگاه دقیق‌تری به این برآوردگر و برآوردگر انقباضی نوع استاین لاسو، ملاحظه می‌شود که اگر $\mathcal{L}_{n,\alpha} \in [0, k_n]$، آنگاه برآوردگر آزمون اولیه بهبودیافته به‌طور یکنواخت بر برآوردگر محدود نشده برتری دارد. از طرف دیگر اگر $\mathcal{L}_{n,\alpha} \in (k_n, \infty]$، آنگاه این برآوردگر مشابه برآوردگر آزمون اولیه لاسو عمل خواهد کرد، بدین معنی که همواره بر برآوردگر محدودنشده برتری نخواهد داشت. در این حالت هر چه از فرضیه صفر فاصله گرفته شود، آنگاه مخاطره برآوردگر آزمون اولیه بهبودیافته به مخاطره برآوردگر محدودنشده همگرا خواهد شد.

[۱۲] Shrinkage Stein-type LASSO Estimator (SSLE)
[۱۳] Improved Preliminary Test LASSO Estimator (IPTLE)

۳ توزیع مجانبی برآوردگرها

در این بخش، رفتار مجانبی برآوردگرهای ارائه شده از منظر توزیعی مورد بررسی قرار داده شده است. در ادامه، برای اثبات روابط، شرایط نظم زیر مفروض می باشند. اگر $n \to \infty$، آنگاه

(i) ماتریس معین مثبت C وجود دارد به‌طوری‌که $n^{-1}C_n \to C$، $C = (C_{ij}), i, j = 1, \dots, n$.

(ii) $\max_{1 \le i \le n}(n\boldsymbol{x}_i^T C_n^{-1}\boldsymbol{x}_i) \to 0$.

فرضیه مقابل $\mathcal{H}_o$ می‌تواند به‌صورت $H\boldsymbol{\beta} = \boldsymbol{h} + \boldsymbol{\delta}$ (فرضیه جایگزین ثابت) در نظر گرفته شود، که در آن $\boldsymbol{\delta} = (\delta_1, \dots, \delta_q)^T$ یک بردار ثابت است. تحت برقراری این فرضیه، اگر $\boldsymbol{\beta}_n^*$ یک برآوردگر دلخواه باشد، توزیع مجانبی $\sqrt{n}(\boldsymbol{\beta}_n^* - \boldsymbol{\beta})$ وقتی $n \to \infty$، با توزیع $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ برابر خواهد بود (صالح، ۲۰۰۶). بنابراین، در این حالت هنگامی که حجم نمونه زیاد است مقایسه برآوردگرها بی‌معنی خواهد بود. لذا، این ایده به ذهن می‌رسد که در حالت مجانبی، نباید یک فرضیه جایگزین ثابت در نظر گرفت. برای به‌دست آوردن نتایج منطقی، به عنوان یک راه‌حل، می‌توان کلاس فرضیه‌های مقابل موضعی را به‌صورت

$$\mathcal{K}_n : H\boldsymbol{\beta} = \boldsymbol{h} + \frac{\boldsymbol{\delta}}{\sqrt{n}}$$

در نظر گرفت. از آن‌جا که انتخاب $\boldsymbol{\delta} = 0$، فرضیه صفر را نتیجه می‌دهد، می‌توان (۲) را حالت خاصی از کلاس $\mathcal{K}_n$ دانست.

فرض کنید در کلاس $\{\mathcal{K}_n\}$، $\sqrt{n}(\boldsymbol{\beta}_n^* - \boldsymbol{\beta})$ دارای توزیع حدی

$$F(\boldsymbol{y}) = \lim_{n\to\infty} P\left(\sqrt{n}(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}) \le y\right)$$

باشد که تابع توزیع مجانبی برآوردگر $\boldsymbol{\beta}_n^*$ نامیده می‌شود و بر اساس این تابع، می‌توان ماتریس پراکندگی Γ^* را به‌صورت

$$\Gamma^* = \int \boldsymbol{y}\boldsymbol{y}^T dF(\boldsymbol{y})$$

تعریف کرد. لازم به ذکر است که تابع مخاطره مجانبی، برابر اثر ماتریس Γ^* می‌باشد و برابر است با

$$R(\hat{\boldsymbol{\beta}}_n^*) = \mathrm{tr}(\Gamma^*) = \lim_{n\to\infty} E\left[n(\boldsymbol{\beta}_n^* - \boldsymbol{\beta})^T(\boldsymbol{\beta}_n^* - \boldsymbol{\beta})\right]. \qquad (۸)$$

لم ۱. (تیل، ۱۹۷۱) فرض کنید مولفه‌های ϵ مستقل و هم‌توزیع با میانگین صفر و واریانس متناهی σ^2 باشند. اگر مولفه‌های X به‌طور یکنواخت کراندار باشند و شرط اول نظم نیز برقرار باشد، آنگاه

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}_p(0, \sigma^2 C^{-1}),$$

که در آن نماد $\xrightarrow{\mathcal{D}}$ همگرایی در توزیع را نشان می‌دهد.

لم ۲. (اشمیت، ۱۹۷۶) در کلاس $\{\mathcal{K}_n\}$ با در نظر گرفتن شرایط نظم، آماره آزمون نسبت درستنمایی $\mathcal{L}_n$ در توزیع به

$$\mathcal{L} = \frac{(HW+\boldsymbol{\delta})^T(HC^{-1}H^T)^{-1}(HW+\boldsymbol{\delta})}{\sigma^2}$$

با توزیع کی‌دو با q درجه آزادی و پارامتر غیرمرکزی

$$\Delta^2 = \sigma^{-2}\boldsymbol{\delta}^T(HC^{-1}H^T)^{-1}\boldsymbol{\delta} = \sigma^{-2}\boldsymbol{\xi}^T C\boldsymbol{\xi} \tag{9}$$

همگرا است، که در آن $W \sim \mathcal{N}_p(0, \sigma^2 C^{-1})$ و $\boldsymbol{\xi} = C^{-1}H^T(HC^{-1}H^T)^{-1}\boldsymbol{\delta}$.

قضیه ۱. با در نظر گرفتن شرایط نظم، در کلاس $\{\mathcal{K}_n\}$، وقتی $n \to \infty$، آن‌گاه

(الف)

$$\hat{\boldsymbol{\beta}}_n^{RL} \xrightarrow{\mathcal{P}} \text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi})) - C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi})) - \boldsymbol{h})$$

(ب)

$$\hat{\boldsymbol{\beta}}_n^{UL} - \hat{\boldsymbol{\beta}}_n^{RL} \xrightarrow{\mathcal{P}} C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi})) - \boldsymbol{h})$$

(ج)

$$\hat{\boldsymbol{\beta}}_n^{PL} \xrightarrow{\mathcal{P}} \text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi})) - C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi})) - \boldsymbol{h})I(\mathcal{L} < \mathcal{L}_\alpha)$$

که در آن $\mathcal{L}_\alpha$ چندک αام بالایی توزیع کی‌دو با q درجه آزادی و پارامتر غیر مرکزی Δ^2 (رابطه ۹) است.

(د)

$$\hat{\boldsymbol{\beta}}_n^{SL} \xrightarrow{\mathcal{D}} \text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi})) - kC^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi})) - h)\mathcal{L}^{-1}$$

(ه)

$$\hat{\beta}_n^{IPL} \quad \xrightarrow{\mathcal{D}} \quad \text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi}))$$

$$-C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi}))-\boldsymbol{h})I(\mathcal{L}<\mathcal{L}_\alpha)$$
$$-k\mathcal{L}^{-1}I(\mathcal{L}>\mathcal{L}_{n,\alpha})C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}(Z)-\boldsymbol{h}),$$

که

$$Z(\boldsymbol{\phi})=(\boldsymbol{\phi}-\boldsymbol{\beta})^TC(\boldsymbol{\phi}-\boldsymbol{\beta})+\lambda_0\sum_{j=1}^{p}|\phi_j|,$$

و در آن $\lambda_n/n\to\lambda_0\geq 0$، $k=q-2$ و $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_p)^T$.

برهان. بر اساس نایت و فو (۲۰۰۰) برآوردگر لاسو در احتمال به کمینه‌کننده $Z(\phi)$ همگرا است. قسمت (الف) را می‌توان با استفاده از قضیه اسلاتسکی و رابطه (۳) و هم‌چنین شرط دوم نظم، به‌صورت زیر نوشت:

$$\hat{\boldsymbol{\beta}}_n^{UL}-C_n^{-1}H^T(HC_n^{-1}H^T)^{-1}(H\hat{\boldsymbol{\beta}}_n^{UL}-h)\xrightarrow{\mathcal{P}}$$
$$\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi}))-C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi}))-\boldsymbol{h}).$$

برای اثبات قسمت (ب)، طبق رابطه (۳)، $\hat{\boldsymbol{\beta}}_n^{UL}-\hat{\boldsymbol{\beta}}_n^{RL}=C_nH^T(HC_n^{-1}H^T)^{-1}(H\hat{\boldsymbol{\beta}}_n^{UL}-\boldsymbol{h})$ که طبق قضیه اسلاتسکی و شرط دوم نظم به $C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi}))-\boldsymbol{h})$ همگرا است. برای اثبات قسمت (ج)، با توجه به لم ۲، $I(\mathcal{L}_n\leq\mathcal{L}_{n,\alpha})\xrightarrow{\mathcal{D}}I(\mathcal{L}\leq\mathcal{L}_\alpha)$. طبق رابطه (۴)، قسمت (ب) همین قضیه و هم‌چنین قضیه اسلاتسکی،

$$\hat{\boldsymbol{\beta}}_n^{PL}\xrightarrow{\mathcal{D}}\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi}))-C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{\phi}}(Z(\boldsymbol{\phi}))-\boldsymbol{h})I(\mathcal{L}\leq\mathcal{L}_\alpha).$$

در قسمت (د)،

$$\lim_{n\to\infty}k_n=\lim_{n\to\infty}\frac{(n-p)(q-2)}{(n-p+2)}=q-2=k. \quad (10)$$

بنابراین طبق رابطه (۶)، قسمت (ب) این قضیه و قضیه اسلاتسکی نتیجه به‌دست می‌آید. قسمت (ه)، بر اساس رابطه (۷) و نیز قسمت‌های (ج) و (د) همین قضیه به‌راحتی نتیجه می‌شود. □

قضیه ۲. با برقراری شرایط نظم، در کلاس $\{\mathcal{K}_n\}$ وقتی $n\to\infty$، آن‌گاه

(الف)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{RL}-\boldsymbol{\beta})\xrightarrow{\mathcal{D}}\text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u}))-C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u}))+\boldsymbol{\delta})$$

(ب)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL} - \hat{\boldsymbol{\beta}}_n^{RL}) \xrightarrow{\mathcal{D}} C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) + \boldsymbol{\delta})$$

(ج)

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{PL} - \boldsymbol{\beta}) &\xrightarrow{\mathcal{D}} \text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) \\ &\quad -C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) + \boldsymbol{\delta})I(\mathcal{L} \leq \mathcal{L}_\alpha),\end{aligned}$$

(د)

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{SL} - \boldsymbol{\beta}) &\xrightarrow{\mathcal{D}} \text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) \\ &\quad -kC^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) + \boldsymbol{\delta})\mathcal{L}^{-1},\end{aligned}$$

(ه)

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{IPL} - \boldsymbol{\beta}) &\xrightarrow{\mathcal{D}} \text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) \\ &\quad -C^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) + \boldsymbol{\delta})I(\mathcal{L} \leq \mathcal{L}_\alpha) \\ &\quad -kC^{-1}H^T(HC^{-1}H^T)^{-1}(H\text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) + \boldsymbol{\delta})\mathcal{L}^{-1}I(\mathcal{L} > \mathcal{L}_\alpha).\end{aligned}$$

برهان. طبق قضیه اسلاتسکی و سن و سینگر (۱۹۹۳) و نیز نایت و فو (۲۰۰۰)، برای اثبات قسمت (الف)، طبق معادله (۳) می‌توان نوشت

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{RL} - \boldsymbol{\beta}) &= \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL} - \boldsymbol{\beta}) \\ &\quad -\sqrt{n}C_n^{-1}H^T(HC_n^{-1}H^T)^{-1}[H(\hat{\boldsymbol{\beta}}_n^{UL} - \boldsymbol{\beta}) + (H\boldsymbol{\beta} - \boldsymbol{h})] \\ &= \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL} - \boldsymbol{\beta}) \\ &\quad -C_n^{-1}H^T(HC_n^{-1}H^T)^{-1}[H\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL} - \boldsymbol{\beta}) + \sqrt{n}(H\boldsymbol{\beta} - \boldsymbol{h})] \\ &= \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL} - \boldsymbol{\beta}) \\ &\quad -nC_n^{-1}H^T(nHC_n^{-1}H^T)^{-1}[H\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL} - \boldsymbol{\beta}) + \sqrt{n}(H\boldsymbol{\beta} - \boldsymbol{h})].\end{aligned}$$

با استفاده از قضیه اسلاتسکی نتیجه می‌شود

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{RL} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) - C^{-1}H^T(HC^{-1}H^T)^{-1}[H\text{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u})) + \boldsymbol{\delta}].$$

برای اثبات (ب) می‌توان نوشت

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL}-\hat{\boldsymbol{\beta}}_n^{RL}) &= \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL}-\hat{\boldsymbol{\beta}}_n^{UL}+C_n^{-1}H^T(HC_n^{-1}H^T)^{-1}(H\hat{\boldsymbol{\beta}}_n^{L}-\boldsymbol{h}))\\ &= \sqrt{n}C_n^{-1}H^T(HC_n^{-1}H^T)^{-1}(H\hat{\boldsymbol{\beta}}_n^{UL}-\boldsymbol{h})\\ &= C_n^{-1}H^T(HC_n^{-1}H^T)^{-1}(H\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL}-\boldsymbol{\beta})+\sqrt{n}(H\boldsymbol{\beta}-\boldsymbol{h}))\\ &= C_n^{-1}H^T(HC_n^{-1}H^T)^{-1}(H\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL}-\boldsymbol{\beta})+\boldsymbol{\delta}).\end{aligned}$$

نتیجه بر اساس قضیه اسلاتسکی برقرار است. برای اثبات (ج)، با استفاده از لم ۲ و هم‌چنین قضیه اسلاتسکی، $\sqrt{n}(\hat{\beta}^{PTL}-\beta)$ در توزیع به رابطه زیر همگرا است

$$\mathrm{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u}))-C^{-1}H^T(HC^{-1}H^T)^{-1}(H\mathrm{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u}))+\boldsymbol{\delta})I(\mathcal{L}\leq\mathcal{L}_\alpha).$$

برای اثبات (د)،

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{SL}-\boldsymbol{\beta})=\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL}-\boldsymbol{\beta})-k_n\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL}-\hat{\boldsymbol{\beta}}_n^{RL})\mathcal{L}_n^{-1}.$$

با توجه به قسمت (ب)، $\mathcal{L}_n^{-1}\xrightarrow{\mathcal{D}}\mathcal{L}^{-1}$ ،$\lim_{n\to\infty}k_n=k$ و به‌کار بردن قضیه اسلاتسکی، می‌توان نوشت

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{SL}-\boldsymbol{\beta}) &\xrightarrow{\mathcal{D}} \mathrm{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u}))\\ &\quad -k\left[(C^{-1}H^T(HC^{-1}H^T)^{-1}(H\mathrm{argmin}_{\boldsymbol{u}}(V(\boldsymbol{u}))+\boldsymbol{\delta})\right]\mathcal{L}^{-1}.\end{aligned}$$

سرانجام برای اثبات (ه)،

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{IPL}-\boldsymbol{\beta}) = \sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^{PL}-\boldsymbol{\beta}-k_n\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL}-\hat{\boldsymbol{\beta}}_n^{RL})\mathcal{L}_n^{-1}I(\mathcal{L}_n>\mathcal{L}_{n,\alpha})\right).$$

با استفاده از لم ۲، نتیجه به راحتی به‌دست می‌آید. □

تذکر ۱. با برقراری شرایط نظم، در کلاس $\{\mathcal{K}_n\}$ وقتی $n\to\infty$، به ازای $\lambda_\circ=\circ$، برآوردگر لاسوی محدود شده بر برآوردگر لاسوی محدود نشده برتری دارد.

برهان. اگر $\lambda_\circ=\circ$ و C یک ماتریس معکوس‌پذیر باشد، بدیهی است که $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{UL}-\boldsymbol{\beta})\xrightarrow{\mathcal{D}}W$ و هم‌چنین $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{RL}-\boldsymbol{\beta})\xrightarrow{\mathcal{D}}W-C^{-1}H^T(HC^{-1}H^T)^{-1}(HW+\boldsymbol{\delta})$، که $W\sim\mathcal{N}_p(\circ,\sigma^2C^{-1})$.

طبق تعریف تابع مخاطره مجانبی (رابطه ۸)، می‌توان نوشت

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}_n^{UL}) &= \sigma^{۲}\,\mathrm{tr}(C^{-۱}) \\ R(\hat{\boldsymbol{\beta}}_n^{RL}) &= \sigma^{۲}\,\mathrm{tr}(C^{-۱}) - \mathrm{tr}(C^{-۱}H^T(HC^{-۱}H^T)^{-۱}HC^{-۱}) \\ &\quad +(H\boldsymbol{\beta}-\boldsymbol{h})^T(HC^{-۱}H^T)^{-۱}HC^{-۱}HC^{-۲}H^T(HC^{-۱}H^T)^{-۱}(H\boldsymbol{\beta}-\boldsymbol{h}). \end{aligned}$$

تحت برقرار فرضیه‌ی $\mathcal{H}_o$،

$$R(\hat{\boldsymbol{\beta}}_n^{UL}) - R(\hat{\boldsymbol{\beta}}_n^{RL}) = \mathrm{tr}(C^{-۱}H^T(HC^{-۱}H^T)^{-۱}HC^{-۱}).$$

از آن‌جا که $C^{-۱}H^T(HC^{-۱}H^T)^{-۱}HC^{-۱}$ یک ماتریس معین مثبت است، مثبت بودن عبارت سمت چپ رابطه فوق نتیجه می‌شود و اثبات کامل می‌گردد. □

۴ مطالعه شبیه‌سازی

در این بخش، با استفاده از یک شبیه‌سازی مونت‌کارلو کارآیی نسبی برآوردگرهای پیشنهاد شده مورد بررسی قرار می‌گیرد. یکی از کاربردی‌ترین انتخاب‌های H و h در بحث انتخاب متغیرهاست. گاهی اوقات، متخصص ادعا می‌کند که برخی از متغیرها اثری در مدل رگرسیون ندارند. به عبارت دیگر، اگر بردار $\boldsymbol{\beta}$ را بتوان به‌صورت $(\beta_۱, \beta_۲)^T$ افراز کرد، آن‌گاه $\boldsymbol{\beta}_۲ = ۰$، متغیرهایی را مشخص می‌کند که بر اساس نظر متخصص در پیش‌گویی مدل تاثیری ندارند. در این حالت، ماتریس H، یک ماتریس همانی با مرتبه p خواهد بود و بنابراین $q = p$. بردار h چنین تعریف می‌شود: بدون از دست دادن کلیت مسئله، فرض کنید k مولفه ابتدایی آن برابر با یک و $p-k$ مولفه انتهایی آن برابر با مقدار $\Delta^۲$، میزان انحراف از فرضیه صفر باشد. برای روشن شدن مطلب، فرض کنید $p = ۵$ و k، تعداد مولفه‌های غیر صفر، برابر با سه باشد. در این صورت $\boldsymbol{h} = [۱, ۱, ۱, \Delta^۲, \Delta^۲]^T$. بدیهی است اگر مقدار $\Delta^۲$، صفر باشد، فرضیه $\mathcal{H}_o : H\boldsymbol{\beta} = \boldsymbol{h}$ برقرار می‌شود. شش مقدار برای $\Delta^۲$ به صورت ۰، ۰/۱، ۰/۵، ۱، ۵ و ۱۰ در نظر گرفته می‌شود.

ماتریس X از توزیع نرمال p-متغیره با بردار میانگین $\boldsymbol{\mu} = ۰$ و ماتریس کوواریانس Σ تولید می‌شود. عناصر غیرقطری ماتریس کوواریانس برابر با $r = ۰, ۰/۲, ۰/۹$ در نظر گرفته شده‌اند. فرض کنید $n = ۱۰۰$ و p مقادیر ۱۰، ۱۵ و ۲۰ را بپذیرد. بردار pتایی $\boldsymbol{\beta}$ بدین صورت در نظر گرفته می‌شود که k مولفه ابتدایی آن

مقدار یک می‌پذیرد و سایر مولفه‌های آن، صفر هستند. برای مثال، وقتی $p=۱۰$ و $k=۵$ است، می‌توان بردار $\boldsymbol{\beta}$ را به‌صورت $(۱,۱,۱,۱,۱,۰,۰,۰,۰,۰)^T$ در نظر گرفت. متغیر پاسخ بر اساس مدل

$$y_i=\sum_{i=۱}^{p}X_i\beta_i+e_i,\quad e_i\sim\mathcal{N}(۰,۵^۲)$$

شبیه‌سازی شده‌اند. از استراتژی‌های بیان شده در این مقاله برای برآورد این پارامتر استفاده می‌شود و هر سناریو، ۱۰۰۰ بار تکرار شده است تا مخاطره برآوردگرها به‌دست آیند. کارآیی نسبی به‌صورت

$$RE(\boldsymbol{\beta}^*;\hat{\boldsymbol{\beta}}^{UL})=R(\hat{\boldsymbol{\beta}}^{UL})/R(\boldsymbol{\beta}^*)$$

محاسبه می‌شود که در آن $\boldsymbol{\beta}^*$ یکی از برآوردگرهای پیشنهادی است که کارآیی نسبی آن باید محاسبه شود. هر مقدار بزرگتر از یک نشان‌دهنده برتری برآوردگر پیشنهادی بر برآوردگر لاسوی تیبشیرانی (۱۹۹۶) است. به عبارت دقیق‌تر چنان‌چه مقدار کارآیی بزرگتر از یک باشد، مخاطره برآوردگر پیشنهادی کمتر از مخاطره برآوردگر لاسو است.

در جداول ۱ تا ۳ مقادیر عددی کارآیی نسبی برآوردگرهای ارائه شده نسبت به برآوردگر لاسوی تیبشیرانی (۱۹۹۶) آورده شده است.

همان‌طور که ملاحظه می‌شود برآوردگرهای پیشنهادی لاسوی محدود شده (RL)، آزمون اولیه لاسو (PTL)، نوع استاین لاسو (SSL) و آزمون اولیه بهبود یافته لاسو (IPTL)، همگی بر برآوردگر لاسو، از لحاظ مخاطره کمتر، برتری دارند.

با نگاهی دقیق‌تر به مقادیر جداول می‌توان نتیجه‌گیری‌های زیر را به‌دست آورد.

(۱) هنگامی‌که فرضیه صفر برقرار باشد ($\Delta^۲=۰$)، برآوردگر RL بهتر از سایر برآوردگرها رفتار می‌کند. هر چه از فرضیه صفر فاصله گرفته شود، از کارآیی این برآوردگر کاسته می‌شود.

(۲) برای مقادیر بزرگ $\Delta^۲$ عدد کارآیی نسبی به شدت کاهش می‌یابد. به‌عبارت دقیق‌تر هر چه از فرضیه صفر فاصله گرفته شود میزان برتری برآوردگرهای پیشنهادی به برآوردگر محدودنشده‌ی لاسو به شدت کاهش می‌یابد، حتی در مواردی، برآوردگر لاسوی محدودنشده بهتر از برآوردگرهای پیشنهادی رفتار می‌کند.

(۳) هرچه درجه همبستگی، r بیشتر شود تاثیر دور شدن از فرضیه صفر بر برتری برآوردگرهای پیشنهادی کاهش می‌یابد. به‌عنوان مثال، در حالت $r=۰$، برآوردگر لاسو به ازای $\Delta^۲=۱۰$ بر برآوردگرهای

جدول ۱: مخاطره نسبی برآوردگرها برای $\Delta^۲$های ثابت، $r = ۰$ و مقادیر مختلف p و $k = ۶$

		ULE	RLE	PTLE			SSLE	IPTLE		
$\Delta^۲$	p			۰/۰۱	۰/۰۵	۰/۱		۰/۰۱	۰/۰۵	۰/۱
	۱۰	۱	۱×۱۰^{۳۲}	۱۴/۳۱۴۶	۴/۷۵۵۸	۳/۰۵۱۵	۳/۱۲۷۰	۱۴/۳۱۴۶	۱۱/۵۰۵۵	۸/۹۵۶۲
۰	۱۵	۱	۱۱۳۹۲/۰۸	۸/۸۶۵۸	۴/۰۴۶۶	۲/۶۷۰۳	۴/۲۲۴۰	۸/۸۶۵۸	۱۱/۸۹۱۷	۸/۲۲۶۹
	۲۰	۱	۷۸/۵۲۲۰	۵/۵۴۶۰	۲/۹۲۳۹	۲/۳۲۸۰	۵/۳۴۷۹	۵/۵۴۶۰	۹/۶۷۹۳	۷/۴۵۵۸
	۱۰	۱	۶×۱۰^{۳۰}	۹/۲۳۳۰	۳/۷۳۶۴	۲/۶۸۷۵	۵/۷۹۵۵	۹/۲۳۳۰	۱۳/۱۰۸۳	۸/۸۲۷۳
۰/۱	۱۵	۱	۶۲۵۸/۶۳	۵/۶۸۸۸	۲/۷۵۵۲	۲/۲۰۲۱	۶/۰۲۸۳	۵/۶۸۸۸	۱۰/۴۹۵۲	۷/۶۵۰۹
	۲۰	۱	۴۸/۲۷۲۸	۵/۴۹۸۸	۲/۸۰۳۹	۲/۰۹۷۲	۶/۲۰۱۸	۵/۴۹۸۸	۱۰/۲۳۱۳	۶/۶۴۵۱
	۱۰	۱	۴×۱۰^{۳۰}	۵/۳۲۲۰	۲/۸۸۱۲	۲/۳۲۱۴	۶/۰۲۳۶	۵/۳۲۲۰	۱۰/۹۵۵۸	۸/۰۸۸۳
۰/۵	۱۵	۱	۴۸۶۷/۳۶	۵/۲۲۵۰	۳/۴۶۹۶	۲/۴۴۹۱	۷/۰۶۸۲	۵/۲۲۵۰	۱۳/۲۹۰۲	۷/۹۵۶۱
	۲۰	۱	۱۰/۲۹۴۰	۵/۹۵۶۵	۲/۹۲۰۰	۲/۱۴۸۷	۱/۹۶۵۵	۵/۹۵۶۵	۵/۱۲۶۰	۴/۳۳۱۶
	۱۰	۱	۳×۱۰^{۳۰}	۶/۵۹۰۴	۳/۱۷۴۶	۲/۵۰۲۹	۷/۵۱۵۴	۶/۵۹۰۴	۱۳/۴۸۷۵	۹/۰۶۰۶
۱	۱۵	۱	۲۴۲/۶۳۶۵	۹/۰۷۵۵	۴/۰۸۰۷	۲/۸۰۱۱	۲/۶۶۲۰	۹/۰۷۵۵	۹/۰۲۵۵	۷/۴۰۱۶
	۲۰	۱	۴/۷۶۴۸	۲/۶۴۵۵	۱/۹۸۶۰	۱/۷۱۲۶	۲/۲۴۱۲	۲/۶۴۵۵	۳/۲۳۰۹	۲/۸۲۵۵
	۱۰	۱	$۲/۶۲۹۲ \times ۱۰^{۴}$	۹/۷۴۳۲	۳/۶۰۱۴	۲/۴۷۴۶	۲/۹۶۷۸	۹/۷۴۳۲	۸/۵۸۲۱	۶/۹۷۷۵
۵	۱۵	۱	۱۱۹/۸۷۱۵	۶/۷۸۰۶	۳/۲۹۱۰	۲/۵۶۳۰	۴/۰۶۲۳	۶/۷۸۰۶	۸/۶۹۹۸	۷/۰۹۶۱
	۲۰	۱	۳/۱۱۷۸	۱/۹۴۳۷	۱/۶۰۱۸	۱/۴۲۳۴	۱/۹۴۴۰	۱/۹۴۳۷	۲/۵۳۴۵	۲/۱۷۹۸
	۱۰	۱	۰/۰۲۳۵	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۰/۶۹۳۳	۱/۰۰۰۰	۰/۶۹۳۳	۰/۶۹۳۳
۱۰	۱۵	۱	۰/۰۱۲۱	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۰/۶۶۹۵	۱/۰۰۰۰	۰/۶۶۹۵	۰/۶۶۹۵
	۲۰	۱	۰/۰۰۷۸	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۰/۶۱۴۱	۱/۰۰۰۰	۰/۶۱۴۱	۰/۶۱۴۱

پیشنهادی برتری دارد در صورتی‌که به ازای $r = ۰/۹$ این برتری تغییر می‌کند.

(۴) با زیاد شدن بعد پارامتر، نظم خاصی در میزان برتری برآوردگرهای پیشنهادی مشاهده نمی‌شود. به نظر می‌رسد عوامل متعددی مانند نسبت تعداد پارامترهای غیرصفر به تعداد پارامتر صفر، مقدار $\Delta^۲$، r و $\cdots$ در کارآیی برآوردگرها موثر باشد. آن چه که اهمیت دارد، در تمامی موارد به‌ازای $\Delta^۲$های کوچک (نزدیک فرضیه صفر) برآوردگرهای پیشنهادی بسیار کاراتر از برآوردگر محدود نشده لاسو است.

(۵) سطح معنی‌داری α تاثیر مستقیمی بر میزان برتری برآوردگرهای PTL و IPTL دارد. با افزایش α، کارآیی برآوردگر آزمون اولیه لاسو کاهش می‌یابد اما این نتیجه، در مورد برآوردگر آزمون اولیه بهبودیافته لاسو برقرار نیست. همان‌طور که در متن اشاره شد، کارآیی این برآوردگر علاوه بر سطح معنی‌داری α

جدول ۲: مخاطره نسبی برآوردگرها برای $\Delta^{۲}$های ثابت، $r = ۰/۲$ و مقادیر مختلف p و $k = ۶$

		ULE	RLE	PTLE			SSLE	IPTLE		
$\Delta^{۲}$	p			۰/۰۱	۰/۰۵	۰/۱		۰/۰۱	۰/۰۵	۰/۱
	۱۰	۱	۷×۱۰^{۳۱}	۱۰/۸۶۲۳	۳/۶۸۱۰	۲/۶۲۴۷	۲/۴۹۰۵	۱۰/۸۶۲۳	۸/۴۱۳۸	۷/۰۵۹۳
۰	۱۵	۱	۱۲۰۰۵/۶۷	۸/۷۱۰۵	۳/۶۲۳۶	۲/۴۲۰۱	۴/۳۴۷۷	۸/۷۱۰۵	۱۰/۹۱۸۸	۷/۶۴۶۴
	۲۰	۱	۷۸/۳۹۷۵	۶/۶۴۸۵	۳/۱۵۳۷	۲/۵۶۹۲	۵/۴۰۳۵	۶/۶۴۸۵	۹/۸۴۶۵	۷/۸۸۶۴
	۱۰	۱	۳×۱۰^{۳۰}	۶/۱۱۷۰	۲/۷۱۶۶	۲/۱۱۱۱	۵/۳۹۶۶	۶/۱۱۷۰	۹/۳۴۶۸	۶/۹۹۹۹
۰/۱	۱۵	۱	۵۹۵۴/۷۰	۵/۷۸۸۵	۲/۸۳۰۷	۲/۳۳۷۲	۶/۳۹۹۲	۵/۷۸۸۵	۱۱/۳۱۵۸	۸/۵۳۷۴
	۲۰	۱	۵۰/۵۰۷۰	۴/۷۵۹۷	۲/۷۹۴۳	۲/۱۷۰۹	۶/۴۶۲۰	۴/۷۵۹۷	۹/۸۷۰۲	۶/۸۹۱۰
	۱۰	۱	۲×۱۰^{۳۰}	۵/۳۱۰۲	۲/۵۰۲۶	۲/۰۶۹۶	۵/۸۶۷۴	۵/۳۱۰۲	۹/۴۷۶۷	۷/۲۳۳۵
۰/۵	۱۵	۱	۵۲۳۰/۶۵	۴/۴۹۴۶	۲/۴۱۹۴	۲/۰۰۱۵	۶/۸۰۱۴	۴/۴۹۴۶	۱۰/۴۱۵۱	۷/۵۷۷۹
	۲۰	۱	۹/۹۶۱۴	۳/۶۳۱۵	۲/۴۰۴۳	۱/۹۵۳۲	۲/۰۴۳۱	۳/۶۳۱۵	۴/۲۱۱۳	۳/۸۰۶۳
	۱۰	۱	۱×۱۰^{۳۰}	۴/۵۹۲۹	۲/۵۴۲۴	۲/۰۷۷۴	۶/۵۹۳۵	۴/۵۹۲۹	۱۰/۱۳۵۸	۷/۳۸۹۱
۱	۱۵	۱	۲۴۸/۱۸۸۶	۷/۸۹۵۶	۳/۶۱۷۲	۲/۳۹۵۱	۲/۷۴۱۴	۷/۸۹۵۶	۸/۲۷۷۷	۶/۶۵۲۹
	۲۰	۱	۴/۷۵۴۱	۲/۶۰۷۷	۱/۸۵۴۶	۱/۶۲۶۴	۲/۳۵۳۲	۲/۶۰۷۷	۳/۱۳۴۵	۲/۸۱۱۱
	۱۰	۱	$۲/۶۸۵۹ \times ۱۰^{۴}$	۸/۳۲۳۲	۳/۲۷۳۰	۲/۳۱۸۷	۴/۵۴۴۰	۸/۳۲۳۲	۷/۶۵۸۵	۶/۳۵۱۴
۵	۱۵	۱	۱۲۶/۰۶۴۶	۷/۶۰۶۲	۳/۱۸۹۴	۲/۴۰۵۱	۴/۴۰۵۲	۷/۶۰۶۲	۹/۰۴۲۵	۷/۱۴۷۳
	۲۰	۱	۳/۴۰۰۸	۱/۹۰۱۷	۱/۴۴۱۹	۱/۳۱۳۷	۲/۲۰۷۹	۱/۹۰۱۷	۲/۵۳۳۳	۲/۱۸۸۶
	۱۰	۱	۰/۱۱۹۸	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۰/۸۲۹۴	۱/۰۰۰۰	۰/۸۲۹۴	۰/۸۲۹۴
۱۰	۱۵	۱	۰/۰۸۴۹	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۰/۹۱۵۵	۱/۰۰۰۰	۰/۹۱۵۵	۰/۹۱۵۵
	۲۰	۱	۰/۰۰۸۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۰/۹۳۴۹	۱/۰۰۰۰	۰/۹۳۴۹	۰/۹۳۴۹

به ثابت انقباضی نیز بستگی دارد. بر اساس نتایج به‌دست آمده از جداول، انتخاب مقدار $\alpha = ۰/۰۵$ منجر به بیشترین کارآیی خواهد شد.

۱.۴ مثال واقعی

در این بخش، کارایی برآوردگرهای بهبودیافته‌ی لاسو در یک مثال واقعی بررسی می‌شود. از داده‌های آمار و ارقام ایالات آمریکا[۱۴] بکر و همکاران (۱۹۸۸) استفاده شده است. این مجموعه داده شامل $n = ۵۰$ ایالت آمریکا و $p = ۸$ متغیر بوده که ۷تای آن توضیحی و یکی (امید به زندگی)، متغیر پاسخ است. توصیفی از متغیرها در این پایگاه داده در جدول ۴ آمده است. آماره‌های توصیفی مربوط به متغیرها در جدول ۵ نشان داده

[۱۴] این مجموعه داده با نام state در بسته datasets در نرم‌افزار R قابل دسترسی است.

جدول ۳: مخاطره نسبی برآوردگرها برای $\Delta^{۲}$های ثابت، $r = ۰/۹$ و مقادیر مختلف p و $k = ۶$

$\Delta^{۲}$	p	ULE	RLE	PTLE			SSLE	IPTLE		
				۰/۰۱	۰/۰۵	۰/۱		۰/۰۱	۰/۰۵	۰/۱
	۱۰	۱	۷×۱۰^{۲۸}	۴/۳۲۰۳	۲/۲۸۵۷	۱/۶۳۸۸	۳/۱۹۸۸	۴/۳۲۰۳	۵/۲۵۵۷	۴/۳۹۹۱
۰	۱۵	۱	۶۱۰۵۳/۱۵	۳/۸۷۳۳	۱/۸۳۹۵	۱/۵۷۶۳	۳/۸۹۱۷	۳/۸۷۳۳	۵/۲۱۲۵	۴/۵۹۶۹
	۲۰	۱	۳۹۸/۱۷۶۸	۳/۲۵۳۷	۱/۷۲۸۴	۱/۴۳۴۶	۴/۳۸۷۰	۳/۲۵۳۷	۵/۵۱۳۱	۴/۴۹۴۳
	۱۰	۱	۱×۱۰^{۲۸}	۵/۰۷۳۱	۲/۱۲۹۸	۱/۵۳۶۴	۴/۹۳۷۴	۵/۰۷۳۱	۷/۲۶۸۶	۴/۹۸۹۸
۰/۱	۱۵	۱	۳۰۲۰۸/۵۲	۲/۸۰۶۴	۱/۸۸۵۲	۱/۵۴۷۵	۴/۹۱۰۲	۲/۸۰۶۴	۶/۲۷۴۹	۴/۸۹۹۰
	۲۰	۱	۲۸۷/۸۷۸۳	۲/۳۴۲۸	۱/۶۱۳۱	۱/۳۸۴۸	۵/۳۱۸۱	۲/۳۴۲۸	۶/۳۸۴۰	۵/۰۰۶۶
	۱۰	۱	۸×۱۰^{۲۷}	۴/۰۰۲۱	۲/۰۲۵۵۹	۱/۵۵۹۹	۵/۶۸۳۲	۴/۰۰۲۱	۷/۸۴۶۰	۵/۴۰۸۵
۰/۵	۱۵	۱	۲۳۹۷۰/۲۶	۳/۴۳۱۸	۱/۹۰۰۵	۱/۵۱۰۸	۶/۲۱۶۸	۳/۴۳۱۸	۸/۰۸۹۷	۵/۵۲۴۵
	۲۰	۱	۴۴/۸۶۴۱	۴/۳۲۲۰	۱/۹۳۶۳	۱/۴۸۰۲	۲/۹۶۴۴	۴/۳۲۲۰	۴/۲۵۵۵	۳/۷۰۶۰
	۱۰	۱	۴×۱۰^{۲۷}	۳/۱۷۷۷	۱/۷۶۹۷	۱/۴۲۷۰	۵/۶۵۰۸	۳/۱۷۷۷	۷/۱۶۱۷	۵/۱۱۰۰
۱	۱۵	۱	۱۲۵۶/۱۴۰	۳/۷۸۷۱	۲/۱۷۲۳	۱/۵۸۳۳	۳/۰۰۴۰	۳/۷۸۷۱	۴/۶۴۳۹	۳/۹۶۴۷
	۲۰	۱	۲۲/۵۴۱۸	۲/۴۰۲۰	۱/۴۰۲۴	۱/۲۳۴۶	۲/۹۸۳۵	۲/۴۰۲۰	۳/۳۸۹۰	۳/۰۶۲۵
	۱۰	۱	۱×۱۰^{۵}	۴/۷۹۶۸	۲/۶۹۲۴	۱/۸۰۸۶	۳/۳۲۳۱	۴/۷۹۶۸	۶/۱۳۵۱	۴/۹۵۱۶
۵	۱۵	۱	۶۴۱/۷۶۸۶	۵/۵۱۳۷	۱/۹۹۵۸	۱/۴۱۶۵	۴/۰۰۹۵	۵/۵۱۳۷	۵/۹۶۶۸	۴/۴۱۲۸
	۲۰	۱	۱۵/۵۲۴۳	۱/۵۸۱۲	۱/۱۸۹۵	۱/۱۰۰۴	۲/۹۴۵۹	۱/۵۸۱۲	۳/۰۷۱۰	۲/۷۸۹۶
	۱۰	۱	۰/۱۲۵۷	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۳۱۰	۱/۰۰۰۰	۱/۰۳۱۰	۱/۰۳۱۰
۱۰	۱۵	۱	۰/۰۴۹۲	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۴۹	۱/۰۰۰۰	۱/۰۰۴۹	۱/۰۰۴۹
	۲۰	۱	۰/۰۳۷۴	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۳۱	۱/۰۰۰۰	۱/۰۰۳۱	۱/۰۰۳۱

شده است.

مدل رگرسیون خطی به داده‌ها برازش داده شده است تا متغیر پاسخ را در حضور متغیرهای پیش‌گو بتوان مدل‌بندی کرد. برآوردگرهای لاسوی محدود نشده (ULE) که همان برآوردگر لاسوی تیبشیرانی (۱۹۹۶) است، آزمون اولیه لاسو (PLE)، نوع استاین لاسو (SLE) و آزمون اولیه بهبودیافته لاسو (IPLE) به منظور برآورد پارامترهای رگرسیونی مورد استفاده قرار گرفته‌اند.

از آن‌جایی‌که یکی از چالش‌های استراتژی‌های برآورد پیشنهاد شده، تعیین ماتریس H و بردار $\boldsymbol{h}$ است و نیاز به نظر متخصص یا مطالعات قبلی دارد، در این قسمت به منظور سادگی و دوری از خطای ناشی از انتخاب نادرست این پارامترها، فرض کنید ماتریس محدودیت H یک ماتریس همانی ۷×۷ در نظر گرفته شده است.

به منظور نشان دادن تاثیر درستی یا نادرستی ادعا بر برآوردگرها دو حالت زیر در نظر گرفته می‌شود:

جدول ۴: توصیفی از متغیرهای داده‌های آمار و ارقام ایالات آمریکا

متغیرها	توصیف
Population	جمعیت
Income	درآمد سرانه
Illiteracy	نرخ بی‌سوادی
Murder	نرخ قتل‌نفس به ازای هر ۱۰۰۰۰۰ نفر
HS Grad	درصد فارغ‌التحصیلان راهنمایی
Frost	میانگین تعداد روزها با دمای زیر صفر
Area	مساحت زمین در مایل‌مربع
Life Exp	امید به زندگی

جدول ۵: آماره‌های توصیفی از متغیرهای داده‌های آمار و ارقام ایالات آمریکا

متغیرها	کمینه	چارک اول	میانه	میانگین	چارک سوم	بیشینه
Population	۳۶۵	۱۰۸۰	۲۸۳۸	۴۲۴۶	۴۹۶۸	۲۱۱۹۸
Income	۳۰۹۸	۳۹۹۳	۴۵۱۹	۴۴۳۶	۴۸۱۴	۶۳۱۵
Illiteracy	۰/۵۰۰	۰/۶۲۵	۰/۹۵۰	۱/۱۷۰	۱/۵۷۵	۲/۸۰۰
Murder	۶۷/۹۶	۴/۳۵۰	۶/۸۵۰	۷/۳۷۸	۱۰/۶۷۵	۱۵/۱۰۰
HS.Grad	۳۷/۸۰	۴۸/۰۵	۵۳/۲۵	۵۳/۱۱	۵۹/۱۵	۶۷/۳۰
Frost	۰/۰۰	۶۶/۲۵	۱۱۴/۵۰	۱۰۴/۴۶	۱۳۹/۷۵	۱۸۸/۰۰
Area	۱۰۴۹	۳۶۹۸۵	۵۴۲۷۷	۷۰۷۳۶	۸۱۱۶۳	۵۶۶۴۳۲
Life.Exp	۶۷/۹۶	۰/۶۲۵	۷۰/۶۷	۷۰/۸۸	۷۱/۸۹	۷۳/۶۰

حالت اول فرض کنید $\boldsymbol{h} = (۰, ۰, ۱۰, ۰/۲, ۰/۷, ۰/۰۶, ۰)^T$. در این صورت فرضیه صفر عبارت است از فرضیه $\mathcal{H}_o : \boldsymbol{\beta} = \boldsymbol{h}$ و بدین معنی است که متغیرهای Population، Income و Area برای برازش مدل مناسب نیستند.

حالت دوم فرض کنید $\boldsymbol{h} = (۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰)^T$. در این صورت فرضیه صفر، $\mathcal{H}_o : \boldsymbol{\beta} = \boldsymbol{h}$ است و بدین معنی است که متغیرهای Population، Income و Area برای برازش مدل مناسب نیستند.

هدف از این انتخاب، نشان دادن کارآیی استراتژی‌های برآورد پیشنهاد شده به منظور بهبود برآوردگر لاسو در کنار یک اطلاع پیشین غیرقطعی است.

کارآیی برآوردگرها با استفاده از میانگین خطای ۵-مجموعه‌ای اعتبارسنجی متقابل به‌دست آمده است.

روش اعتبارسنجی متقابل k-زیرمجموعه‌ای، روشی بسیار رایج در یافتن خطای پیش‌گویی یک برآوردگر در مثال‌های واقعی، هنگامی که از مقدار واقعی پارامتر اطلاعی وجود ندارد، می‌باشد. در این روش، مجموعه داده‌ها به k قسمت تقسیم می‌شود. فرآیند بدین‌گونه تعریف می‌شود که در هر مرحله، یکی از این k قسمت که زیرمجموعه آزمون نام دارد کنارگذاشته می‌شود و با استفاده از $k - ۱$ قسمت باقی‌مانده که مجموعه‌ی آموزشی نام دارد پارامتر مجهول، برآورد می‌شود. با استفاده از این برآورد در مجموعه آزمون، متغیر پاسخ، پیش‌گویی شده و اختلاف این مقدار پیش‌بینی‌شده با مقدار واقعی متغیر پاسخ در مجموعه آزمون محاسبه می‌گردد و خطای پیش‌گویی نامیده می‌شود. پس از اینکه از تمامی این k قسمت به‌عنوان مجموعه‌ی آزمون استفاده شد، میانگین این k خطای پیش‌گویی، به‌عنوان خطای پیش‌گویی این فرآیند در نظر گرفته می‌شود. برای اینکه اثر تصادفی بودن در این فرآیند از بین برود، کل این فرآیند، ۱۰۰۰ بار تکرار می‌شود. بر اساس توضیحات بیان شده، قدرت پیش‌گویی نسبی برآوردگر به‌صورت نسبت میانگین خطای پیش‌گویی نسبی برآوردگر دلخواه نسبت به میانگین خطای پیش‌گویی برآوردگر لاسو محاسبه می‌شود.

جدول ۶ شامل مقادیر قدرت پیش‌گویی نسبی سه برآوردگر پیشنهادی آزمون اولیه لاسو، انقباضی نوع استاین لاسو و آزمون اولیه بهبودیافته لاسو نسبت به برآوردگر لاسو به ازای $\alpha = ۰/۱$ است. از آن‌جایی که اعداد جدول ۶ همگی بزرگتر از یک هستند، می‌توان نتیجه گرفت که هر سه برآوردگر پیشنهادی میانگین خطای پیش‌گویی کمتری نسبت به برآوردگر لاسو دارند. اما با نگاهی دقیق‌تر مشاهده می‌شود که در بین همه برآوردگرها، برآوردگر آزمون اولیه بهبودیافته لاسو بهتر از سایر برآوردگرها بردار پارامترهای مدل را پیش‌گویی می‌کند. این برتری در حالت اول بسیار واضح است. با توجه به مقدار قدرت پیش‌گویی نسبی برآوردگر محدودنشده، به نظر می‌رسد این ادعا، صحیح‌تر از ادعای مطرح شده در حالت دوم است و همان‌طور که در متن اشاره شد در نزدیکی فرضیه صفر، برآوردگر آزمون اولیه بهبودیافته بهتر از سایر برآوردگرها رفتار می‌کند. بنابراین پیشنهاد می‌شود در این مثال، به جای برآوردگر لاسو از این برآوردگر در مسئله برآورد و پیش‌گویی استفاده شود. مقادیر برآورد شده برآوردگر لاسوی محدودنشده و آزمون اولیه بهبودیافته لاسو در دو حالت بیان شده در جدول ۷ آمده است.

جدول ۶: جدول خطاهای پیش‌گویی نسبی برآوردگرها در مجموعه داده آمار و ارقام ایالت‌های آمریکا

	RLE	PTLE	SLE	IPLE
حالت اول	۲۲/۲۶۰۸	۱/۰۰۰۰	۱/۰۱۹۳۷	۱/۰۱۹۷۰
حالت دوم	۱/۰۰۰۶	۱/۰۰۰۱	۱/۰۰۰۰۶	۱/۰۰۰۱۴

جدول ۷: جدول مقادیر برآوردگرهای لاسو و آزمون اولیه بهبودیافته لاسو به ازای $\alpha = ۰/۱$

	ULE	IPTLE حالت اول	IPTLE حالت دوم
Population	۰/۰۰۰۰۳	۰/۰۰۰۰۳	۰/۰۰۰۰۳
Income	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
Illiteracy	۰/۰۰۰۰	۰/۱۰۰۴	۰/۰۰۰۰
Murder	−۰/۲۶۳۹	−۰/۲۷۲۶	−۰/۲۶۳۸
HS.Grad	۰/۰۳۷۹	۰/۰۴۸۹	۰/۰۳۷۹
Frost	−۰/۰۰۳۳	−۰/۰۰۳۶	−۰/۰۰۳۳
Area	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰

۵ بحث و نتیجه‌گیری

در این مقاله، با قرار دادن یک محدودیت بر روی مدل و استفاده از تکنیک‌های آزمون اولیه و انقباضی، بهبودی روی برآوردگر لاسو انجام شد. در واقع، برآوردگرهای آزمون اولیه‌ی لاسو (PTLE)، نوع استاین لاسو (SLE) و آزمون اولیه بهبودیافته لاسو (IPTLE) در زیرفضای محدود شده معرفی گردیدند. کارآیی برآوردگرهای پیشنهاد شده در حالتی که حجم نمونه (n) بیشتر از بعد پارامتر (p) است، بررسی شد (جدول‌های ۱–۳). با توجه به نتایج به‌دست آمده از مطالعات عددی، وقتی فرضیه صفر تقریبا برقرار است، در بین همه برآوردگرهای پیشنهاد شده، برآوردگر آزمون اولیه بهبودیافته لاسو بهتر از سایر برآوردگرها عمل می‌کند. از طرفی، هر چه از فرضیه صفر فاصله گرفته شود، هیچ‌یک از برآوردگرهای نوع استاین لاسو و آزمون اولیه لاسو بر دیگری برتری ندارند. برآوردگر آزمون اولیه لاسو با کاهش α از عملکرد بهتری برخوردار است. با توجه به نتایج عددی ارائه شده در این مقاله، می‌توان نتیجه گرفت که با در اختیار داشتن اطلاعات غیرنمونه‌ای، برآوردگرهای بهبودیافته لاسو و

به‌ویژه برآوردگر آزمون اولیه بهبودیافته لاسو بر برآوردگر لاسوی تیبشیرانی از دو منظر کارآیی نسبی و قدرت پیش‌گویی برتری دارد.

تقدیر و تشکر

نویسندگان مقاله از نظرات و پیشنهادات داوران محترم و هیئت تحریریه مجله علوم آماری که باعث ارتقای مقاله شد، کمال تشکر و قدردانی را دارند.

مراجع

Ahmed, S. E. and Nicol, C. J. (2012), An Application of Shrinkage Estimation to the Nonlinear Regression Model, *Computational Statistics and Data Analysis*, **56**, 3309-3321.

Bancroft, T. A. (1944), On Biases in Estimation due to the Use of Preliminary Tests of Significance, *The Annals of Mathematical Statistics*, **15**, 190-204.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988), *The New S Language*, Wadsworth and Brooks/Cole.

Belloni, A. and Chernozhukov, V. (2013), Least Squares after Model Selection in High-Dimensional Sparse Models, *Bernoulli*, **19**, 521-547.

Fan, J. and Li, R. (2001), Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

Gruber, M. (1998), *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*, Marcel Dekker, New York.

Hansen, B. E. (2016), The Risk of James Stein and LASSO Shrinkage, *Econometric Reviews*, **35**, 1456-1470.

Hoerl. A. E., Kennard. R. W. (1970), Ridge Regression Baised Estimation for Non-orthogonal Problems, *Thechnometrics*, **12**, 69-89.

James, W. and Stein, C. (1961), Estimation of Quadratic Loss, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 361-379.

Judge, G. G. and Bock, M. E. (1978), *The Statistical Implication of Pre-test and Stein-rule Estimators in Econometrics*, North-Holland, New York.

Kibria, B. M. G. and Saleh, A. K. M. E. (2003), Effect of W, LR, and LM Tests on the Performance of Preliminary Test Ridge Regression Estimators, *Journal of the Japan Statistical Society*, **33**, 119-136.

Kibria, B. M. G. and Saleh, A. K. M. E. (2012), Improving the Estimators of the Parameters of a Probit Regression Model: A Ridge Regression Approach, *Journal of Statistical Planning and Inference*, **142**, 1421-1435.

Knight, K. and Fu, W. (2000), Asymptotics for LASSO-type Estimators, *Annals of Statistics*, **28**, 1356-1378.

Lian, H. (2013), Shrinkage Estimation and Selection for Multiple Functional Regression, *Statistica Sinica*, **23**, 51-74.

Raheem, S. M. and Ahmed, S. E. (2011), Positive-Shrinkage and Pretest Estimation in Multiple Regression: a Monte Carlo Study with Applications, *Journal of Iranian Statistical Society*, **10**, 267-289.

Saleh, A. K. M. E. (2006), *Theory of Preliminary Test and Stein-Type Estimation with Applications*, John Wiley, New York.

Saleh, A. K. M. E., Arashi, M. and Tabatabaey, S. M. M. (2014), *Statistical Inference for Models with Multivarite t-Distributed Errors*, John Wiley, New Jersy.

Saleh, A. K. M. E. and Kibria, B. M. G. (2011), On Some Ridge Regression Estimators: a Nonparametric Approach, *Journal of Nonparametric Statistics*, **23**, 819 - 851.

Schmidt, P. (1976), *Econometrics*, Mercel Dekker, New York.

Sclove, S. L., Morris, C. and Radhakrishnan, R. (1972), Non-optimality of Preliminary Test Estimators of the Mean of a Multivariate Normal Distribution, *The Annals of Mathematical Statistics*, **43**, 1481-1490.

Sen, P. K. and Singer, J. M. (1993), *Large Sample Methods in Statistics: An Introduction with applications*, Chapman and Hall, New York.

Sengupta, D. and Jammalamadaka, S. R. (2003), *Linear Models: An Integrated Approach*, World Scientific Publishing Company, Singapore.

Shalabh (1998), Improved Estimation in Measurement Error Models through Stein-rule Procedure, *Journal of Multivariate Analysis*, **67**, 35-48.

Shalabh (2001), Pitman Closeness Comparison of Least Squares and Stein-rule Estimators in Linear Regression Models with Non-normal Disturbances, *The American Journal of Mathematical and Management Sciences*, **21**, 89-100.

Stein, C. (1956), Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 197-206.

Theil, H. (1971), *Principles of Econometrics*, John Wiley, New York.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the LASSO, *Journal of the Royal Statistical Society*, Series B, **58**, 267-288.

Tikhonov, A. N. (1963), Solution of Incorrectly Formulated Problems and the Regularization Method, *Doklady Akademii Nauk SSSR*, **151**, 501-504, Translated in *Soviet Mathematics*, **4**, 1035-1038.

Touloumis, A. (2015), Nonparametric Stein-type Shrinkage Covariance Matrix Estimators in High Dimensional Setting, *Computational Statistics and Data Analysis*, **83**, 251-261.

Zou, H. (2006), The Adaptive Lasso and its Oracle Properties, *Journal of the American Statistical Association*, **101**, 1418-1429.

Zou, H. and Hastie, T. (2005), Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society*, Series B, **67**, 301-320.

Studying Limiting Behaviour of Shrinkage Estimators in Penalized Regression Model with Rectangular Norm

M. Norouzirad, M. Arashi

Department of Statistics, Shahrood University of Technology, Shahrood, Iran.

Abstract: Penalized estimators for estimating regression parameters have been considered by many authors for many decades. Penalized regression with rectangular norm is one of the mainly used since it does variable selection and estimating parameters, simultaneously. In this paper, we propose some new estimators by employing uncertain prior information on parameters. Superiority of the proposed shrinkage estimators over the least absoluate and shrinkage operator (LASSO) estimator is demonstrated via a Monte Carlo study. The prediction rate of the proposed estimators compared to the LASSO estimator is also studied in the US State Facts and Figures dataset.

Keywords: Asymptotic distribution, Improved estimator, LASSO estimator, Prediction error, Rectangular norm, Stein-type shrinkage estimator.

Mathematics Subject Classification (2010): 62F12, 62J07, 62F30.