مروری بر مشکلات متداول رگرسیون و راه‌حل‌های آن‌ها

زهرا جعفریان مورکانی [۱]، حیدرعلی مردانی فرد✉ [۲]

تاریخ دریافت: ۱۴۰۱/۰۷/۲۰
تاریخ پذیرش: ۱۴۰۱/۱۲/۰۲

چکیده:
مدل رگرسیون خطی معمولی به‌صورت $Y = X\beta + \varepsilon$ است و برآورد پارامتر β عبارت است از: $\hat{\beta} = (X'X)^{-1}X'Y$. بااین‌حال در هنگام استفاده از این برآوردگر به‌صورت عملی، ممکن است مشکلات خاصی مانند مشکل انتخاب متغیر، هم‌خطی، مدل با ابعاد بالا، کاهش بعد و وجود خطای اندازه‌گیری به وجود آید که استفاده از برآوردگر $\hat{\beta}$ را مشکل می‌سازد. در اغلب این مشکلات، مسئله اصلی عدم معکوس‌پذیری ماتریس $X'X$ است. برای رفع آن‌ها راه‌حل‌های متعددی ارائه شده است. در این مقاله ضمن مروری بر این مشکلات، مجموعه‌ای از راه‌حل‌های معمول و متداول و همچنین چند روش خاص و پیشرفته (که کمتر مورد اقبال همگان است ولی بااین‌حال توانایی بالقوه‌ای در رفع هوشمند این مشکلات دارند) را بررسی می‌کنیم.

واژه‌های کلیدی: هم‌خطی، کاهش بعد، خطای اندازه‌گیری، رگرسیون ستیغی، شبه‌برآورد، رگرسیون معکوس قطعه‌قطعه شده

۱ مقدمه

یکی از متداول‌ترین و پرکاربردترین روش‌های آماری در تحلیل داده‌ها رگرسیون است که در حالت معمول آن، اثر یک یا چند متغیر روی یک متغیر دیگر به‌صورت یک رابطه عموماً خطی سنجیده می‌شود. در مدل رگرسیون خطی، رابطه بین متغیر وابسته Y و متغیرهای مستقل $X_0, X_1, \cdots, X_p$ به‌صورت زیر در نظر گرفته می‌شود:

$$y = \mathbf{X}'\boldsymbol{\beta} + \varepsilon, \qquad (۱)$$

که در آن بردار $\boldsymbol{\beta}$ بردار ضرایب رگرسیونی، $\mathbf{X} = (X_0, X_1, \cdots, X_p)'$ بردار متغیرهای مستقل با $X_0 \equiv 1$ بوده و ε خطای مدل است. وقتی متغیرها را روی یک نمونه n تایی اندازه‌گیری کرده باشیم مدل به‌صورت ماتریسی زیر نوشته می‌شود:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (۲)$$

که در آن X ماتریس داده‌های متغیرهای مستقل است و بردار $\mathbf{y}$ مقادیر متغیر وابسته (پاسخ) و بردار $\boldsymbol{\varepsilon}$ هم بردار خطاهای تصادفی است. به‌طورمعمول برآورد $\boldsymbol{\beta}$ با روش کمترین مربعات (که در صورت نرمال و مستقل بودن خطاها معادل با برآورد حداکثر درستنمایی است) عبارت است از:

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}, \qquad (۳)$$

بااین‌حال، در استفاده از برآورد فوق لازم است به برخی پیش‌فرض‌های مهم دقت کرد. هنگام استفاده از داده‌های همبسته یا داده‌های با ابعاد بالا در رگرسیون خطی با معکوس ناپذیری ماتریس $X'X$ مواجه می‌شویم که محاسبه $\hat{\boldsymbol{\beta}}$ را غیرممکن و یا آن را بسیار ناپایدار می‌کند به‌گونه‌ای که با اندک تغییر جزئی در داده‌ها مقدار برآوردگر به‌طور قابل‌ملاحظه‌ای تغییر می‌کند. اخیراً، لی و یین [۸] پیشنهاد کردند که خطاهایی به پیش‌بینی کننده‌ها با ساختار کوواریانس شناخته شده افزوده شود و برآورد حاصل را «شبه برآورد» نامیدند. واگر و همکاران [۱۲] نشان دادند که در حالتی خاص، داده‌های دارای خطا، ارتباط تنگاتنگی با برآوردگر معروف ستیغی دارد. یک روش جدید انتخاب متغیر بر اساس *شبه‌فاصله اطمینان* می‌باشد که برای پیش‌بینی کننده‌های همبسته و مشکل "p بزرگ، n کوچک" پیشنهاد می‌شود. در این مقاله، روش شبه برآورد

[۱] دانشجوی کارشناسی ارشد رشته آمار، گروه ریاضی، دانشگاه یاسوج
✉ [۲] دانشگاه یاسوج، گروه ریاضی (نویسنده مسئول: h_mardanifard@yu.ac.ir)

پیشنهادی و انتخاب متغیر بر اساس آن‌هم مورد مطالعه قرار می‌گیرد.

۲ ابزارها و مفاهیم مفید رگرسیون

۱.۲ ضریب تعیین و آماره C_p مالو

ضریب تعیین ($R^۲$) کسری از کل تغییرات متغیر پاسخ است که توسط متغیرهای مستقل تبیین می‌شود. مقدار این ضریب همواره بین ۰ و ۱ است و اگر به یک نزدیک باشد، نشان‌دهنده‌ی برازش بهتر خط رگرسیون است و اگر به صفر نزدیک‌تر باشد، خط رگرسیون به‌خوبی به داده‌ها برازش داده نشده است. معادله ضریب تعیین به‌صورت زیر است:

$$R^۲ = \frac{SSR}{SST} = ۱ - \frac{SSE}{SST}$$

اگر در مقایسه مدل‌ها تعداد پارامترها برابر نباشد از ضریب تعیین تعدیل‌شده استفاده می‌شود که در آن اثر تعداد متغیرها تعدیل‌شده است:

$$R^۲_{adj} = ۱ - \frac{n-۱}{n-p} \times \frac{SSE}{SST} = ۱ - \frac{n-۱}{n-p}\left(۱ - R^۲\right)$$

علاوه بر این‌ها، آماره C_p مالو[۳] برای اولین بار توسط مالو [۹] به‌صورت زیر تعریف شد:

$$C_p = \frac{MSE_p}{MSE} - (n - ۲p)$$

که در آن $MSE = \frac{SSE}{n-p}$ و SSE_p از یک مدل با $k = p - ۱$ متغیر به وجود آمده است. در اینجا k و p نشان‌دهنده تعداد متغیرهای مستقل و تعداد پارامترها در مدل می‌باشد. امید ریاضی این آماره عبارت است از: $E(C_p) = p$ و در استفاده از این آماره متغیری برای ورود به مدل مناسب‌تر است که C_p کمتری دارد [۲].

۲.۲ معیار اطلاع آکائیکه

معیارهای اطلاع مانند معیار اطلاع آکائیکه[۴] (به‌اختصار AIC) و معیار اطلاع بیزی[۵] (به‌اختصار BIC) از روی مقدار بهینه تابع درستنمایی تعریف می‌شود که در مدل‌های رگرسیونی با توزیع خطای نرمال می‌توان به‌صورت زیر از روی مجموع مربعات خطا محاسبه کرد[۵]:

$$AIC = \frac{۱}{n}(SSE + ۲d\hat{\sigma}^۲)$$
$$BIC = \frac{۱}{n}(SEE + d\ln(n)\hat{\sigma}^۲)$$

طبیعی است که هر چه مقدار این آماره‌ها کمتر باشد مناسبت مدل بیشتر است.

۳.۲ رگرسیون تاوانیده

روش حداقل مربعات معمولی هنگامی‌که تعداد متغیرها بیشتر از نمونه باشد عملکرد ناامیدکننده‌ای دارد. رگرسیون تاوانیده یا روش‌های انقباضی، به این صورت است که مدل رگرسیونی جریمه می‌شود و اعمال این تاوان به متغیرها اجازه می‌دهد تا ضریبی نزدیک به صفر یا مساوی صفر داشته باشند. توجه کنید که انقباض نیاز به انتخاب یک پارامتر تنظیم‌کننده[۶] (مثلاً λ) دارد که تعیین‌کننده‌ی میزان انقباض است و با روش اعتبارسنجی متقابل تعیین می‌شود[۴]. در رگرسیون تاوانیده با تابع تاوان $P_\lambda(\boldsymbol{\beta})$ برآورد پارامترها از روی بهینه‌سازی تابع زیر به دست می‌آید:

$$Q_\lambda(\boldsymbol{\beta}) := SSE(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}),$$

که در آن $SSE(\boldsymbol{\beta}) = \sum_i (y_i - \beta_۰ - \sum_j \beta_j x_{ij})^۲$.

۱.۳.۲ رگرسیون ستیغی

رگرسیون ستیغی[۷] برای رسیدن به پیش‌بینی بهتر در برخورد با هم‌خطی، توسط هورل و کنارد[۴]، پیشنهاد شد. در این روش مقدار تاوان به‌صورت $P_\lambda(\boldsymbol{\beta}) = \sum \beta_j^۲$ است و در نتیجه برآورد ستیغی با مینیمم کردن عبارت زیر به دست می‌آید:

$$SSE(\boldsymbol{\beta}) + \lambda \sum_{j=۱}^{p} \beta_j^۲.$$

در مسائل بهینه‌سازی، این موضوع معادل است با مینیمم کردن $SSE(\boldsymbol{\beta})$ با قید $\sum \beta_j^۲ = c$. در سمت چپ شکل ۱ نمودار سطح مقطع و ناحیه قید بهینه‌سازی فوق رسم شده است.

[3] Mallows' C_p
[4] Akaike information criterion
[5] Bayesian information criterion
[6] Tuning parameter
[7] Ridge regression

نتیجه این بهینه‌سازی، برآورد زیر را به دست می‌دهد:

$$\hat{\boldsymbol{\beta}}^{(R)} = (X'X + \lambda I)^{-1} X'\mathbf{y}.$$

۲.۳.۲ رگرسیون لاسو

رگرسیون لاسو[8] برای اولین بار توسط تیبشیرانی [۱۱] پیشنهاد شد که انتخاب متغیر و برآورد پارامتر به‌طور هم‌زمان توسط آن، انجام می‌گیرد. در این روش تابع تاوان به‌صورت مجموع قدر مطلق ضرایب تعریف می‌شود و در نتیجه باید عبارت زیر برای برآورد ضرایب مینیمم شود:

$$SSE(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

مشابه رگرسیون ستیغی، در اینجا هم مینیمم سازی فوق معادل با مینیمم کردن تابع $SSE(\beta)$ با قید $\sum |\beta_j| = c$ است. نمودار سطح مقطع تابع هدف و همچنین ناحیه قید این بهینه‌سازی در سمت راست شکل ۱ رسم شده است.

۳.۳.۲ مقایسه رگرسیون ستیغی و لاسو

در هر دو روش ستیغی و لاسو، محل اولین برخورد سطح مقطع‌ها با ناحیه محدودیت، جواب رگرسیون تاوانیده می‌باشد و ناحیه محدودیت مربوط به رگرسیون ستیغی دایره‌ای شکل و نقاط روی دایره نسبت به هم ارجحیتی ندارند؛ بنابراین، احتمال برخورد بیضی و ناحیه روی دایره در یک نقطه مشخص (مانند نقاط روی محورها) صفر است و هرچند که ممکن است برآورد ضرایب رگرسیون ستیغی به صفر خیلی نزدیک باشند ولی دقیقاً صفر نمی‌شود. در طرف مقابل، ناحیه محدودیت مربوط به لاسو دارای تیزی و گوشه می‌باشد، بنابراین ناحیه شرطی و سطح مقطع‌ها با احتمال مثبت در محورها و نقاط گوشه برخورد داشته و ممکن است در ابعاد بالاتر، تعداد بیشتری از ضرایب صفر شوند.

۴.۲ رگرسیون مؤلفه‌های اصلی

روش تحلیل مؤلفه‌هاي اصلی براي اولین بار در سال ۱۹۰۱ توسط کارل پیرسون[9] دانشمند انگلیسی بیان شد که شامل تجزیه مقادیر ویژه ماتریس کواریانس است [۳]. این روش برای داده‌ها با حجم نمونه زیاد مورداستفاده قرار می‌گیرد. اولین جهت مؤلفه اصلی داده‌ها همان جهتی است که بیشترین تغییرات مشاهدات در آن وجود دارند و باعث تعریف خطی می‌شود که تا حد امکان به داده‌ها نزدیک می‌باشد. مدل رگرسیونی (۱) را در نظر بگیرید که در آن بردار تصادفی $\mathbf{X}$ دارای امید ریاضی μ و ماتریس واریانس‌ـ کواریانس Σ است. هدف این است که بجای بردار $\mathbf{X}$ بتوان از بردار

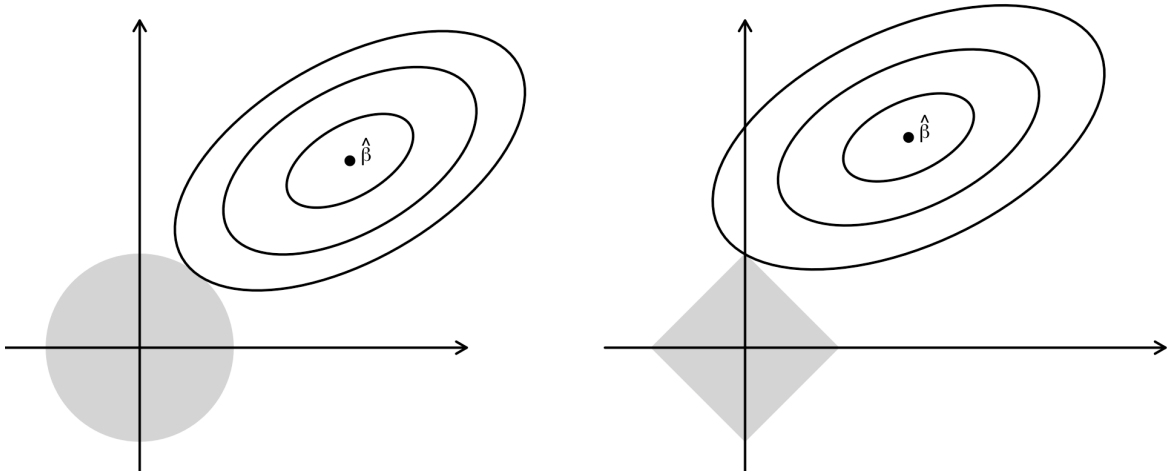
$$\mathbf{X}^* = (X_1^*, \ldots, X_q^*)'$$

استفاده کرد به‌طوری‌که $q \ll p$ باشد درحالی‌که $\mathbf{X}^*$ نسبت به $\mathbf{X}$ اطلاعات زیادی از دست ندهد. برای یافتن چنین $\mathbf{X}^*$ی راه‌های زیادی وجود دارد که یکی از این روش‌ها تجزیه طیفی ماتریس واریانس‌ـ کوواریانس است. تجزیه طیفی به این صورت عمل می‌کند که باید Σ به‌صورت $P\Lambda P'$ نمایش داده شود. اگر مقادیر λ_jها از بزرگ به کوچک مرتب شوند، عملاً X_j^*ها به ترتیب بیشترین تا کمترین واریانس مرتب می‌شوند. به‌این‌ترتیب می‌توان از q تا X_j^* ابتدایی، یعنی q تا بردار ویژه متناسب با q تا بزرگ‌ترین مقادیر ویژه استفاده کرد.

۵.۲ رگرسیون معکوس قطعه‌قطعه شده

برآورد ماتریس واریانس‌ـ کواریانس با استفاده از قطعه‌بندی دامنه‌ی متغیر پاسخ، برای اولین بار توسط لی[10] [۷] مطرح شد. این روش، مشهورترین روش کاهش بعد در رگرسیون است و برای کاهش بعد متغیر مستقل $\mathbf{X}$ بدون انجام هیچ‌گونه فرآیند برازش مدل پارامتری یا ناپارامتری ارائه می‌شود که در عوض رگرسیون متغیر پاسخ تک متغیره y در برابر $\mathbf{X}$ چند متغیره اجرا می‌گردد. هدف رگرسیون معکوس قطعه‌قطعه شده[11] (به‌اختصار SIR)، کشف یک تصویر از متغیر کمکی pـبعدی $\mathbf{X}$، روی یک زیرفضای خطی k بعدی است که حاوی اکثر اطلاعات در خصوص پاسخ y است. هر پارامتر β که در زیرفضای k بعدی وجود داشته باشد، به‌عنوان «جهت» کاهش بعد مؤثر[12] (به‌اختصار e.d.r.) بیان و کوچک‌ترین e.d.r. زیرفضای مرکزی نامیده می‌شود که دارای کمترین بعد و بیشترین واریانس است. لی [۶] تشریح کرد هنگامی‌که y تغییر می‌کند،

[8] LASSO
[9] Karl Pearson
[10] Li
[11] Sliced Inverse Regression
[12] effective dimension reduction



شکل ۱: مقایسه نمودار سطح مقطع‌های رگرسیون ستیغی (چپ) و لاسو (راست)

رابطه $E(\mathbf{X}|y)$ موجب ترسیم یک منحنی می‌شود که در R^p، منحنی رگرسیون معکوس نام می‌گیرد. درصورتی‌که متغیرهای کمکی $\mathbf{X}$ طوری استانداردشده باشند که میانگین آن بردار صفر و ماتریس واریانس‌ـ کواریانس آن ماتریس همانی باشد، آنگاه این زیرفضا با فضای e.d.r. منطبق و سبب شناسایی جهت‌های اصلی تغییرات می‌شود. الگوریتم رگرسیون معکوس قطعه‌قطعه شده به‌وسیله‌ی لی [۶] به‌صورت زیر مطرح شد:

۱. ابتدا X به‌صورت زیر استاندارد می‌شود:

$$\tilde{\mathbf{x}}_i = \hat{\Sigma}_x^{-\frac{1}{2}}(\mathbf{x}_i - \bar{\mathbf{x}}), \qquad i = 1, \ldots, n,$$

که در آن $\hat{\Sigma}_x$ و $\bar{\mathbf{x}}$ به ترتیب ماتریس واریانس‌ـ کوواریانس نمونه و میانگین نمونه X هستند.

۲. محدوده y، به H قطعه، یعنی $I_1, ..., I_H$ به‌صورت صعودی تقسیم می‌گردد و نسبتی از y_iها که در قطعه hام قرار می‌گیرد $\hat{P}_h$ نامیده می‌شود؛ یعنی $\hat{P}_h = n^{-1}\sum \delta_h(y_i)$, می‌باشد که در آن $\delta_h(y_i)$ مقادیر ۰ یا ۱ را بسته به اینکه آیا y_i در قطعه h قرار گرفته است یا نه، به خود می‌گیرد.

۳. در هر قطعه، میانگین نمونه $\tilde{\mathbf{x}}_i$ها که با $\hat{\mathbf{m}}_h$ نشان داده می‌شود محاسبه می‌شود:

$$\hat{\mathbf{m}}_h = \frac{1}{n}\hat{P}_h \sum \tilde{\mathbf{x}}_i = \frac{1}{n_h}\sum \tilde{\mathbf{x}}_i,$$

که در آن n_h تعداد مشاهدات در هر قطعه می‌باشد. در واقع، $\hat{\mathbf{m}}_h$ برآوردی از $E(Z|y)$ در هر قطعه هستند.

۴. ماتریس کوواریانس وزنی $\hat{V} = \Sigma_{h=1}^{H}\hat{P}_h\hat{\mathbf{m}}_h\hat{\mathbf{m}}_h'$ تشکیل‌شده و سپس مقادیر ویژه و بردارهای ویژه $\hat{V}$ محاسبه می‌شود. در واقع $\hat{V}$ برآوردی از $cov(\hat{E}(Z|y))$ می‌باشد.

۵. قرار می‌دهیم:

$$\hat{\beta}_k = \hat{\eta}_k\hat{\Sigma}_x^{-\frac{1}{2}}, \; k = 1, \ldots, K$$

که در آن $\hat{\eta}_k$ها بردارهای ویژه متناظر با K مقدار ویژه بزرگ ماتریس $\hat{V}$ هستند.

۶.۲ شبه برآورد

ایده شبه برآورد توسط لی و یین [۸] برای مدل‌های رگرسیونی با خطای اندازه‌گیری ارائه شد. هنگامی‌که اندازه نمونه کمتر از تعداد متغیرهای پیش‌بین است، به علت معکوس‌پذیر نبودن ماتریس $(X'X)$، روش رگرسیون معمولی درست عمل نمی‌کند و در این حالت، یکی از روش‌های مناسب استفاده از شبه برآورد است [۱۰]. در مدل‌های معمول رگرسیونی با خطای اندازه‌گیری فرض می‌شود که بردار $q \times 1$ بُعدی پیش‌بینی کننده $\mathbf{W}$ با پیش‌بین‌های واقعی $\mathbf{X}$ از طریق یک معادله خطی مانند زیر مرتبط است:

$$\mathbf{W} = \gamma + \Gamma\mathbf{X} + \boldsymbol{\delta}, \qquad (۴)$$

که در آن γ و Γ به ترتیب، بردار غیر تصادفی $q \times 1$ بُعدی و ماتریس غیر تصادفی $p \times q$ بُعدی بوده و $\boldsymbol{\delta}$ یک بردار خطای تصادفی با توزیع نرمال چند متغیره مستقل از $\mathbf{X}$ و $\mathbf{Y}$ می‌باشد. شبه برآورد هنگامی مطرح می‌شود که متغیر مستقل، دارای خطای اندازه‌گیری باشد و برآوردگر به‌دست‌آمده به‌عنوان «شبه برآورد» نام می‌گیرد که

ساده‌ترین حالت معادله (۴) با مقادیر $p = q$ و $\Gamma = I_p$ است؛ یعنی $\mathbf{W} = \mathbf{X} + \boldsymbol{\delta}$.

معمولاً، یک نمونه کمکی وجود دارد که اطلاعاتی در مورد رابطه بین پیش‌بینی کننده اصلی $\mathbf{X}$ و پیش‌بینی کننده جایگزین $\mathbf{W}$ مانند $\Sigma_{wx} = cov(\mathbf{W}, \mathbf{X})$ ارائه می‌کند. با استفاده از این برآورد کوواریانس، می‌توان از پیش‌بینی کننده جایگزین $\mathbf{W}$ استفاده کرد که تا حد امکان با پیش‌بینی کننده واقعی $\mathbf{X}$ هماهنگ شود. بدین ترتیب، این روش در بعضی مواقع ناچار می‌باشد به متغیرهای پیش‌بین خطایی بیفزاید. مثالی از متغیرهای با خطای اندازه‌گیری، داده‌های فشارخون است که معمولاً خود داده‌ها، واقعی نیستند و تقریبی هستند.

۱.۶.۲ شبه برآورد و انتخاب متغیر

برآورد پارامترها در روش شبه برآورد به دو صورت زیر است:

$$\hat{\boldsymbol{\beta}}^{pseudo} = (X'X + n\Sigma_\delta)^{-1}X'\mathbf{y} \qquad (۵)$$

$$\tilde{\boldsymbol{\beta}}^{pseudo} = (W'W)^{-1}W'\mathbf{y} \qquad (۶)$$

در روش شبه برآورد، هنگامی‌که از رابطه (۵) استفاده می‌شود، نگرانی به وجود نمی‌آید زیرا برای هر دو حالت $n > p$ و $n < p$ درست عمل می‌کند اما رابطه دوم یعنی (۶)، زمانی که $n < p$ است، جواب مطلوبی ندارد و برای استفاده از این معادله وقتی $Z = (X, \mathbf{y})$ کل نمونه‌های مشاهده‌شده باشد، یک نمونه جایگزین ترکیبی به‌صورت $Z^* = (W^*, \mathbf{y}^*)$ به دست می‌آید که $\mathbf{y}^*$ تکرار بردار $\mathbf{y}$ به‌اندازه m بار می‌باشد و ماتریس W^*، m بار با استفاده از معادله (۴) تولید می‌شود. تعداد تکرارها یعنی m هم طوری پیدا می‌شود که اندازه نمونه جدید $(m \times n)$ بزرگ‌تر یا مساوی تعداد متغیرها باشد. به‌طور دقیق‌تر:

$$\mathbf{y}^* = (\mathbf{y}', \ldots, \mathbf{y}')',$$

$$W^* = (W_1', \ldots, W_m')',$$

$$W_i = X + \boldsymbol{\delta}_i.$$

می‌توان ثابت نمود که برآورد به‌دست‌آمده با استفاده از شبه مشاهدات (یعنی $\tilde{\boldsymbol{\beta}}^{pseudo}$) وقتی تعداد تکرارها خیلی زیاد شود (وقتی $m \to \infty$) به برآورد $\hat{\boldsymbol{\beta}}^{pseudo}$ میل می‌کند.

۷.۲ شبه‌فاصله اطمینان

ازآنجاکه می‌توان برای یک Σ_δ چندین شبه نمونه ایجاد کرد، می‌توان چندین شبه برآورد به دست آورد و توزیع نمونه‌برداری از شبه برآورد مستقیماً تقریب زده و منجر به یک روش انتخاب متغیر جدید برای شبه برآوردها می‌شود. فرض کنید تولید شبه مشاهدات N بار انجام شود و برای هر $p = ۱, ۲, \ldots,$ قرار دهید:

$$\bar{\beta}_j = N^{-۱}\sum_{k=۱}^{N}\hat{\beta}_{kj},$$

$$S.E.(\bar{\beta}_j) = \frac{\sum_{k=۱}^{N}(\hat{\beta}_{kj} - \bar{\beta}_j)^۲}{N - ۱},$$

حال مشابه تئوری بوت استرپ[۱۳]، می‌توان برای β_j فاصله اطمینانی به‌صورت زیر ساخت:

$$\beta_j \quad \in \quad \left(\bar{\beta}_j \pm Z_{\frac{\alpha}{۲}} \times S.E.(\bar{\beta}_j)\right)$$

این فاصله اطمینان را اصطلاحاً یک شبه‌فاصله اطمینان[۱۴] (به‌اختصار PCI) می‌نامیم زیرا با استفاده از شبه مشاهدات به دست می‌آید. حال هرکدام از β_jهایی که فاصله اطمینان آن‌ها شامل عدد صفر می‌شود در مدل تأثیرگذار نیستند، بنابراین صفر در نظر گرفته می‌شوند. این روش انتخاب متغیر اصطلاحاً S-Pseudo نام دارد. بر اساس PCI، با این روش می‌توان یک شبه برآورد *پراکنده‌شده*[۱۵] به دست آورد.

۱.۷.۲ سطح مقطع‌ها

در هندسه به شکلی که با بریدن یک شیء سه‌بعدی توسط شیء دوبعدی به وجود می‌آید، اصطلاحاً سطح مقطع گفته می‌شود و دانستن سطح مقطع‌های تابع معادل با دانستن تابع است زیرا ارتباط یک‌به‌یک بین آن‌ها برقرار می‌باشد. اگر تابع ما، یک فرم درجه دوم با ماتریسی مانند Σ باشد، هر Σ سطح مقطع‌های خاص خود را دارد. در واقع در فرم‌های درجه دوم اگر ماتریس Σ معین مثبت و متقارن باشد، سطح مقطع‌های $x'\Sigma x$ به فرم بیضی هستند اما ممکن است مورب، افقی یا قائم باشند که دانستن اینکه بیضی‌ها کجا و به چه شکل هستند معادل با دانستن ماتریسΣ است.

جهت‌های اصلی و فرعی بیضی‌هایی که سطح مقطع‌های $x'\Sigma x$ هستند، همان بردارهای ویژه Σ بوده و اندازه قطرهای بیضی هم

[13] Bootstrap

[14] Pseudo Confidence Interval

[15] Sparsed pseudo-estimation

متناسب با λ_i یا همان مقادیر ویژه می‌باشند و به این خاطر است که مقادیر ویژه و بردارهای ویژه برای شناخت Σ بکار گرفته می‌شود.

۸.۲ جهت‌های اصلی هسین

این روش بر اساس یافتن جهت‌های اصلی ماتریس هسین[۱۶] (به‌اختصار phd) است که نخستین بار توسط لی [۷] بیان شده است. ماتریس هسین با H(x) نمایش داده می‌شود. فرض کنید ماتریس A به‌صورت $A = \bar{H}(x)\Sigma_x$ تعریف‌شده باشد. به کمک روش تجزیه طیفی این ماتریس به‌صورت زیر درمی‌آید:

$$A = \sum \lambda_j \mathbf{e}_i \mathbf{e}_i' = P\Lambda P'$$

که در آن Λ ماتریسی قطری متشکل از مقادیر ویژه‌ی A روی قطر اصلی آن بوده و ستون‌های ماتریس P متشکل از بردارهای ویژه‌ی A است:

$$\Lambda = \mathbf{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_p\right),$$
$$P = \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_p \end{pmatrix}.$$

کاهش بعد در تجزیه طیفی به این صورت است که λ_i ها مرتب می‌شوند و از جایی به بعد حذف می‌گردند. به‌بیان‌دیگر، بجای Λ از

$$\Lambda^* = \mathbf{diag}\left(\lambda_1, \ldots, \lambda_q, 0, \ldots, 0\right)$$

استفاده می‌شود که بعضی از مقادیر ویژه صفر شده‌اند. درواقع، $P\Lambda^* P'$ به وجود می‌آید. دقت کنید که رابطه مقادیر و بردار ویژه برای ماتریس A به‌صورت $A\mathbf{e}_i = \lambda_i \mathbf{e}_i$ است اما ازآنجاکه هر بردار در راستای قطرهای بیضی دارای خاصیت $A\mathbf{e}_i = \lambda_i \mathbf{e}_i$ می‌باشد، باید محدودیت $\mathbf{b}'\Sigma\mathbf{b} = 1$ اعمال شود که ماتریس هسین آن به‌صورت $\bar{H}(x) = E(H(x))$ است.

۳ مشکلات معمول در رگرسیون

۱.۳ انتخاب متغیر

هدف از انتخاب متغیر در رگرسیون، یافتن بهترین مجموعه پیش‌بینی کننده‌ها از میان بسیاری از متغیرها برای گنجاندن در یک مدل است. رگرسیون گام‌به‌گام با بهره‌گیری از معیارهای ضریب تبیین تعدیل‌شده، آماره C_p مالو، معیار اطلاع آکائیکه و بیزی برای این کار می‌تواند استفاده شود. همچنین از روش‌های شبه برآورد و رگرسیون تاوانیده لاسو هم می‌توان برای انتخاب متغیرهای مهم استفاده نمود.

۲.۳ هم‌خطی

به وجود ارتباط خطی بین متغیرهاي پیش‌بین در مدل رگرسیون خطی چندگانه اصطلاحاً هم‌خطی گفته می‌شود. مشکل هم‌خطی باعث کاهش دقت برآورد ضرایب رگرسیون شده و خطای استاندارد $\hat{\beta}_j$ را زیاد می‌کند به‌نحوی‌که عمده ضرایب متغیرهای واقعاً مهم هم معنادار نمی‌شوند. برای مواجهه با هم‌خطی می‌توان از روش‌های شبه برآورد، رگرسیون مؤلفه‌های اصلی و رگرسیون تاوانیده‌ی ستیغی استفاده نمود. این روش‌ها، یا با ترکیب متغیرهای پیش‌بین همبسته، تعداد کمتری متغیر پیش‌بین ناهمبسته می‌سازند و یا اساساً با انقباض پارامترهای مدل جلوی ناپایداری و زیادشدن خطای استاندارد آن‌ها را می‌گیرند.

۳.۳ مدل با ابعاد بالا

اخیراً، در سال ۱۹۹۰ استفاده از داده‌هایی که در آن تعداد متغیرهاي پیش‌بین (p) بسیار بیشتر از تعداد مشاهدات (n) می‌باشد، توسط افرون و هستی [۳]، آغاز شد. افزایش زیاد ابعاد داده‌ها یک مشکل اساسی در مبحث داده‌کاوي هم هست. اصطلاح *مدل با ابعاد بالا*[۱۷] مشخصاً زمانی به کار می‌رود که در یک نمونه، تعداد پارامترهای نامعلوم، بیشتر از اندازه نمونه باشد ($n \ll p$). نخستین بار ایده روش برآورد کننده ستیغی آستانه[۱۸] توسط شاو و دنگ [۱۰] ارائه شد. در این روش، پیشنهاد می‌گردد که برآوردگر رگرسیون ستیغی، زمانی که بردار ضرایب دارای مقادیر کوچک است، آستانه‌گذاری شود. مدل خطی $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ را در نظر بگیرید. در برخی از مواقع، نیازی به برآورد خود بردار $\boldsymbol{\beta}$ نیست و هدف پیش‌بینی یا برآورد متغیر پاسخ است؛ بنابراین، تصویر $\boldsymbol{\beta}$ روی $R(X)$ در نظر گرفته می‌شود که به‌صورت $\boldsymbol{\theta} = X'(X'X)^{-}X\boldsymbol{\beta} = QQ'\boldsymbol{\beta}$ می‌باشد. توجه کنید که همواره $\boldsymbol{\theta} \in R(X)$ بوده و $\boldsymbol{\theta} = \boldsymbol{\beta}$ است اگر و فقط اگر $\boldsymbol{\beta} \in R(X)$ باشد، علاوه بر این همواره داریم: $X\boldsymbol{\theta} = X\boldsymbol{\beta}$ و مدل اول را می‌توان به‌صورت $\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ هم نوشت که در این حالت برآورد $\boldsymbol{\theta}$ برای

[16] Principal Hessian Directions
[17] High-dimesional model
[18] Treshhold ridge method

استنباط در مورد پارامترهای $X\boldsymbol{\theta} = X\boldsymbol{\beta}$ و پیش‌بینی $\mathbf{y}$ کافی است. درحالی‌که β ممکن است حاوی تعداد زیادی مؤلفه صفر باشد، $\boldsymbol{\theta}$ می‌تواند هیچ مؤلفه صفری نداشته باشد، اگرچه بسیاری از اجزای $\boldsymbol{\theta}$ ممکن است نزدیک به صفر باشند. برآوردگر رگرسیون ستیغی آستانه‌ای به این صورت تاوانیده می‌شود که اگر قدر مطلق هر یک از مؤلفه‌های $\hat{\theta}$ از یک حد آستانه کمتر شود آنگاه آن ضریب برابر صفر قرار داده می‌شود و در غیر این صورت مقدار آن ضریب به $\tilde{\theta}$ منتقل می‌شود:

$$\tilde{\theta}_j = \hat{\theta}_j I_{|\hat{\theta}_j|>a_n}, \qquad j = ۱, \ldots, p.$$

تعیین مقدار آستانه به‌صورت زیر انجام می‌شود:

$$a_n = Cn^{-\alpha}, \qquad ۰ < \alpha \leq \frac{۱}{۲}, \quad C > ۰,$$

و مقادیر دلخواه α و C به n بستگی ندارند. از این آستانه می‌توان به‌عنوان یک روش انتخاب هم متغیر استفاده نمود. روش‌هاي شبه برآورد، روش‌های کاهش ابعاد از جمله SIR و PHD، رگرسیون تاوانیده از جمله رگرسیون ستیغی و لاسو، رگرسیون مؤلفه‌های اصلی (PCA) می‌توانند در مدل‌های با ابعاد بالا استفاده شوند. دقت کنید که روش‌های کلاسیک از جمله رگرسیون بهترین زیرمجموعه با استفاده از معیارهای مجموع توان‌هاي دوم خطا، C_p مالو و AIC وBIC و... در این‌گونه مدل‌ها کارایی ندارند زیرا برآورد پارامترها در این روش‌ها وابسته به معکوس ماتریس $X'X$ است که در این مدل‌ها معکوس‌پذیر نیست.

۴.۳ کاهش بعد

هدف از کاهش بعد، برآورد جهت بردارهای ستون X یا فضای ستونی X است که فضای کاهش ابعاد مجموعه‌ای از پیش‌بینی کننده‌های مهم از میان تمام ترکیبات خطی X به دست می‌دهد. فرض کنید $y = g(\mathbf{X}, \varepsilon)$ باشد که $\mathbf{X}$ تشکیل یک فضای p بُعدی می‌دهد. اکنون، هدف ساختن زیر فضای k بُعدی از طریق ترکیبات خطی $\mathbf{X}$ است که $y = g(\beta_۱'\mathbf{X}, \ldots \beta_k'\mathbf{X}, \varepsilon)$ تقریباً همان فضای قبلی را تولید کند. اکنون، برای کاهش بعد فرض کنید $E(y|\mathbf{X}) = f(\mathbf{X}) = h(\beta_۱'\mathbf{X}, \ldots, \beta_k'\mathbf{X}) = h(W_۱, \ldots, W_k)$. هدف یافتن β_jها است به‌طوری‌که کمترین خطای ممکن را داشته باشیم. از روش‌های زیر برای کاهش بعد می‌توان استفاده کرد:

- رگرسیون مؤلفه اصلی (PCR).
- رگرسیون معکوس قطعه‌قطعه شده (SIR).
- جهت‌های اصلی هسین (PHD).

۵.۳ وجود خطاي اندازه‌گیری

برای نخستین بار ادکاک[۱]، مدل‌های رگرسیونی با خطای اندازه‌گیری را مطرح کرد. مدلی که در آن متغیر پیش‌بین یا متغیر پاسخ یا هر دو با خطا اندازه‌گیري شوند را مدل رگرسیونی با خطاي اندازه‌گیري می‌گویند و اگر متغیر پیش‌بین Xدارای خطاي اندازه‌گیري باشد، در عمل به‌جای مشاهده x_i متغیر تصادفی $w_i = x_i + \delta_i$ مشاهده می‌شود که در آن x_i متغیر مشاهده نشده، w_i متغیر مشاهده‌شده و δ_i متغیري تصادفی است که بیان‌کننده خطاي اندازه‌گیري است. در مدل با خطاي اندازه‌گیري با توجه به رابطه $var(W) = var(X)+var(\delta) > var(X)$، واریانس متغیر پیش‌بین قابل‌مشاهده W از واریانس متغیر پیش‌بین مشاهده نشده X بزرگ‌تر است. در مواجهه با مدل با خطای اندازه‌گیری می‌توان از روش‌های شبه برآورد و رگرسیون‌های تاوانیده استفاده کرد.

۴ بحث و نتیجه‌گیری

مدل های رگرسیون خطی ممکن است با چالش های متعددی در کاربردهای عملی روبرو شوند، از جمله انتخاب متغیر، همبستگی، بعد بالا و خطای اندازه گیری. این مسائل می‌توانند استفاده از برآوردگر سنتی را به چالش بکشانند. در این مقاله، یک مجموعه جامع از روش‌های متداول و پیشرفته برای حل این مشکلات ارائه شده است. استفاده از این روش‌ها می‌تواند در رفع این چالش‌ها موثر باشد.

مراجع

[1] Adock, R.J. (1877). Note on the method of least squares. *Analyst*, **4**, 183-184.

[2] Chowdhury, M.Z.I. and Turin, T.C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health*, **8(1)**, e000262.

[3] Hastie, T., Tibshirani, R., and Friedman, J. (2001).*The elements of statistical learning*. Springer series in statistics. New York, NY, USA.

[4] Horel, A.E. (1962). Applications of ridge analysis to regression problems. *Chem. Eng. Progress.*, **58**, 54-59.

[5] Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research*, **33(2)**, 188–229.

[6] Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86(414)**, 316-327.

[7] Li, K.C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, **87(420)**, 1025-1039.

[8] Li, B. and Yin, X. (2007). On surrogate dimension reduction for measurement error regression: an invariance law. *The Annals of Statistics*, **35(5)**, 2143-2172.

[9] Mallows, C.L. (1973). Some Comments on Cp. *Technometrics*, **15(4)**, 661–675.

[10] Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, **40(2)**, 812-831.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58(1)**, 267-288.

[12] Wager, S., Wang, S., and Liang, P.S. (2013). Dropout training as adaptive regularization. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, **1**, 351–359.

On common linear regression problems and their solutions

ZahraJafarian Moorakani[1] and Heydar Ali Mardani-Fard✉ [2]

Abstract:

The traditional linear regression model is represented as $Y = X\beta + \varepsilon$, with the estimated parameter β calculated as $\hat{\beta} = (X'X)^{-1}X'Y$. However, when implementing this estimator in practical applications, several issues may arise, such as variable selection, collinearity, high dimensionality, dimension reduction, and measurement error, which can make it challenging to use the above estimator. The primary problem in most of these cases is the singularity of the matrix $X'X$. A variety of solutions have been proposed to address these problems. In this article, we review these issues and present a comprehensive set of common solutions, as well as some advanced and less commonly used methods, that have the potential to address these problems in an intelligent manner.

Keywords: Colliniarity, dimension reduction, measurement error, ridge regression, psuedo estimation, SIR regression.

[1]**Yasouj University, Yasouj, IRAN**

[2]✉**Yasouj University, Yasouj, IRAN (corresponding author: h_mardanifard@yu.ac.ir)**