

برآورد استوار نسبت به مشاهددهای دورافتاده در رگرسیون خطی در حضور هم خطی چندگانه

سارا جذن^۱، سید مرتضی امینی^۲

تاریخ دریافت: ۱۳۹۵/۱۲/۲۰

تاریخ پذیرش: ۱۳۹۶/۶/۲۶

چکیده:

یکی از عوامل تأثیرگذار در تحلیل آماری داده‌ها، وجود مشاهددهای دورافتاده است. به روش‌هایی که تحت تأثیر مشاهددهای دورافتاده قرار نمی‌گیرند، روش‌های آماری استوار گفته می‌شود. علاوه بر وجود مشاهددهای دورافتاده، وجود وابستگی خطی میان متغیرهای پیشگو، که از آن با عنوان هم خطی چندگانه یاد می‌شود و نیز تعداد زیاد متغیرها در مقابل اندازه کم نمونه، به خصوص در مدل‌های تنک با بعد بالا، از دیگر مشکلاتی هستند که منجر به کاهش کارایی استنباط‌های حاصل از روش‌های کلاسیک رگرسیونی می‌شوند.

در این مقاله، ابتدا معایب روش رگرسیونی کلاسیک کمترین توان‌های دوم در مقابل مشاهددهای دورافتاده، هم خطی چندگانه و مدل‌های تنک را بررسی می‌کنیم. سپس به معرفی و بررسی روش‌های رگرسیون استوار و رگرسیون توانانیده به عنوان راهکارهای حل این مشکلات می‌پردازیم. همچنین با در نظر گرفتن مشاهددهای دورافتاده و هم خطی چندگانه و یا مدل‌های تنک به طور هم‌زمان به بررسی روش‌های رگرسیون استوار توانانیده می‌پردازیم.

در نهایت به منظور مقایسه عملکرد برآوردگرهای مختلف مطرح شده در این مقاله، ابتدا سه مطالعه شبیه‌سازی را انجام داده و سپس به تحلیل یک مجموعه داده واقعی با استفاده از روش‌های رگرسیون استوار توانانیده می‌پردازیم.

واژه‌های کلیدی: مشاهددهای دورافتاده، رگرسیون استوار، هم خطی چندگانه، مدل تنک، رگرسیون توانانیده.

۱ مقدمه

مشاهددهای دورافتاده در داده‌ها است، که از این تأثیر به عنوان عدم استواری برآوردگر یاد می‌شود. لذا یافتن روش‌هایی که تحت تأثیر این نوع مشاهددهای قرار نگیرند امری ضروری به حساب می‌آید، که این روش‌ها را رگرسیون استوار می‌نامند. از جمله روش‌های رگرسیون استوار می‌توان به برآوردگر رگرسیونی کمترین قدر مطلق انحرافات [۴]، M -برآوردگرها [۹]، کمترین میانه توان‌های دوم [۱۴]، کمترین توان‌های دوم پیراسته [۱۵] و S -برآوردگرها [۱۷] اشاره کرد.

علاوه بر وجود مشاهددهای دورافتاده، وجود وابستگی خطی

در تحقیقات مختلف با مسائلی سر و کار داریم که در آنها با استفاده از مجموعه‌ای از متغیرهای مستقل به پیشگویی رفتار مجموعه‌ای دیگر از متغیرها می‌پردازیم. یکی از روش‌های آماری که کاربرد وسیعی در این گونه مسائل دارد، رگرسیون خطی چندگانه است. اولین و معروف‌ترین برآورد رگرسیون خطی منسوب به گاوس و لژاندر است، که به روش کمترین توان‌های دوم معروف است. یکی از عواملی که برآوردگر کمترین توان‌های دوم را تحت تأثیر قرار می‌دهد وجود

^۱ دانش‌آموخته کارشناسی ارشد آمار، دانشگاه تهران، ایران

^۲ عضو هیئت علمی گروه آمار، دانشگاه تهران، ایران

۲ مشاهددهای دورافتاده و روش‌های رگرسیون استوار

مدل رگرسیونی خطی به صورت

$$y = X\beta + \varepsilon, \quad (1)$$

را در نظر بگیرید. که در آن y بردار $n \times 1$ مشاهددهای متغیر پاسخ،

$$X = [X_1 \dots X_p] = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$$

ماتریس $n \times p$ مشاهددهای متغیرهای پیشگو، β بردار $p \times 1$ ضرایب نامعلوم رگرسیونی و ε بردار $n \times 1$ خطاها با میانگین صفر و ماتریس واریانس کواریانس $\sigma^2 I$ است. هدف برآورد بردار پارامترهای نامعلوم β از مجموعه مشاهددهای نمونه (y, X) است. یکی از متداول‌ترین روش‌های برآورد بردار β روش کمترین توان‌های دوم است. برای مدل رگرسیونی خطی (؟؟) برآوردگر رگرسیونی کمترین توان‌های دوم^۳ (LS) به صورت

$$\hat{\beta}_{LS} = \arg \min_{\beta} \sum_{i=1}^n (y_i - x_i \beta)^2. \quad (2)$$

به دست می‌آید. رابطه (؟؟) به آسانی برآوردگر

$$\hat{\beta}_{LS} = (X'X)^{-1} X'y. \quad (3)$$

را نتیجه می‌دهد. یکی از عواملی که برآوردهای رگرسیونی کمترین توان‌های دوم را تحت تأثیر قرار می‌دهد وجود مشاهددهای دورافتاده^۴ در داده‌ها است. دو نوع اصلی مشاهددهای دورافتاده مشاهددهای دورافتاده عمودی و نقاط نافذ هستند که به صورت زیر تعریف می‌گردند. مشاهده (y_k, x_k) که در آن مقدار y_k دور از خط رگرسیونی‌ای که اکثر نقاط مجموعه داده از آن تبعیت می‌کنند قرار دارد یک مشاهده دورافتاده عمودی نامیده می‌شود. مشاهده (y_k, x_k) که در آن x_k دور از مجموعه اکثریت x_i های مشاهده شده در

میان متغیرهای پیشگو، که از آن با عنوان هم‌خطی چندگانه یاد می‌شود، تعداد زیاد متغیرها در مقابل اندازه کم نمونه و مدل‌های تنک با بعد بالا، از دیگر مسائلی هستند که منجر به بی‌اعتباری برآوردهای حاصل از روش کمترین توان‌های دوم می‌شوند. راهکار منظم سازی و یا رگرسیون توان‌یافته روشی کلی برای فائق آمدن بر چنین مسائلی است. از جمله روش‌های رگرسیون توان‌یافته می‌توان به روش رگرسیون ستیغی، روش کمترین قدر مطلق انقباض و عملگر انتخاب و روش تور مرتجع اشاره کرد.

طرح کلی این مقاله به این صورت است. در بخش دوم، پس از بیان مقدمات رگرسیون خطی و روش کلاسیک کمترین توان‌های دوم، به معرفی مشاهددهای دورافتاده و نقاط نافذ و تأثیرات آنها بر این برآوردگر کلاسیک رگرسیونی می‌پردازیم. سپس انواع برآوردهای استوار را معرفی می‌کنیم. در بخش سوم، پس از معرفی هم‌خطی چندگانه، مشکل تعداد زیاد متغیرها در مقابل اندازه کم نمونه و مدل‌های تنک و بررسی تأثیرات این مشکلات بر روش‌های رگرسیونی کلاسیک، به معرفی روش‌های رگرسیونی توان‌یافته می‌پردازیم. در بخش چهارم، روش رگرسیون استوار کمترین توان‌های دوم پیراسته توان‌یافته بر اساس سه نوع متداول تابع تاوان را به عنوان راهکاری برای مقابله هم‌زمان با مشکلات هم‌خطی چندگانه و یا مدل‌های تنک و مشاهددهای دورافتاده معرفی می‌کنیم. در بخش پنجم، روش رگرسیون استوار کمترین قدر مطلق انحرافات توان‌یافته بر اساس سه نوع متداول تابع تاوان را به عنوان راهکاری دیگر برای مقابله هم‌زمان با مشکلات هم‌خطی چندگانه و یا مدل‌های تنک و مشاهددهای دورافتاده بررسی می‌کنیم. در بخش ششم، به منظور مقایسه عملکرد برآوردهای مختلف مطرح شده در این مقاله، ابتدا سه مطالعه شبیه‌سازی را انجام داده و سپس به تحلیل یک مجموعه داده واقعی با استفاده از روش‌های رگرسیون استوار توان‌یافته می‌پردازیم.

^۳ Least Squares

^۴ Outliers

M-برآوردگرها با

$$\hat{\beta}_M = \arg \min_{\beta} \sum_{i=1}^n \rho(y_i - \mathbf{x}_i \beta),$$

که در آن $\rho(\cdot)$ تابعی محدب، متقارن حول صفر و با مینیمم یکتا در صفر است، برآوردگر کمترین میانه توان‌های دوم^۹ (LMS) با

$$\hat{\beta}_{LMS} = \arg \min_{\beta} \text{Med}_{1 \leq i \leq n} (y_i - \mathbf{x}_i \beta)^2,$$

برآوردگر کمترین توان‌های دوم پیراسته^{۱۰} (LTS) با

$$\hat{\beta}_{LTS} = \arg \min_{\beta} \sum_{i=1}^h (y - \mathbf{X}\beta)_{i:n}^2, \quad (۴)$$

که در آن

$$(y - \mathbf{X}\beta)_{1:n}^2 \leq \dots \leq (y - \mathbf{X}\beta)_{n:n}^2$$

مرتب شده توان دوم مانده‌ها و

$$h = \left\lfloor \frac{n}{\gamma} \right\rfloor + \left\lfloor \frac{p+1}{\gamma} \right\rfloor, \quad h = \lfloor \gamma \delta n \rfloor \quad (۵)$$

مقادیر توصیه شده‌ای برای h هستند و S -برآوردگرها با

$$\hat{\beta}_S = \arg \min_{\beta} S(y_1 - \mathbf{x}_1 \beta, \dots, y_n - \mathbf{x}_n \beta),$$

که در آن $s = S(y_1 - \mathbf{x}_1 \beta, \dots, y_n - \mathbf{x}_n \beta)$ از معادله

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\frac{y_i - \mathbf{x}_i \beta}{s}\right) = K$$

به دست می‌آید. که K اغلب برابر با $E(\rho(Z))$ در نظر گرفته می‌شود که Z دارای توزیع نرمال استاندارد است و ρ تابعی با شرایط خاص است، اشاره کرد.

به علت عدم استواری برآوردگر رگرسیونی LAD در مقابل نقاط نافذ نقطه فروریزش مجانبی این برآوردگر به بالاتر از

نمونه قرار دارد یک نقطه نافذ^۵ نامیده می‌شود. یکی از راه‌های تشخیص مشاهده‌های دورافتاده نمودارهای جزئی نقاط نافذ^۶ است که به صورت زیر تعریف می‌گردد.

فرض کنید $\mathbf{X}_{[k]}$ ماتریس $n \times (p-1)$ حاصل از حذف k امین ستون $\mathbf{X}$ و $\mathbf{u}_k$ و $\mathbf{v}_k$ به ترتیب مانده‌های حاصل از رگرسیون y روی $\mathbf{X}_{[k]}$ و رگرسیون $\mathbf{X}_k$ روی $\mathbf{X}_{[k]}$ باشند. در این صورت k امین ضریب رگرسیونی y روی $\mathbf{X}$ می‌تواند از رگرسیون ساده $\mathbf{u}_k$ روی $\mathbf{v}_k$ تعیین شود. نمودار جزئی نقاط نافذ برای $\hat{\beta}_k$ به صورت نمودار پراکندگی $\mathbf{u}_k$ در برابر $\mathbf{v}_k$ به همراه خط رگرسیونی ساده برازش داده شده به آنها تعریف می‌شود. به این ترتیب می‌توان مشاهده‌های دورافتاده بالقوه را در این نمودارها تشخیص داد.

در مواجهه با مشاهده‌های دورافتاده معمولاً استفاده از روش‌های رگرسیونی استوار^۷ (نیرومند) توصیه می‌شود که به شیوه‌هایی اطلاق می‌شوند که نه تنها در مقابل مشاهده‌های دورافتاده مقاوم هستند، بلکه منجر به برآوردگرهایی با کارایی نسبتاً مطلوب نیز می‌شوند.

یکی از متداول‌ترین ملاک‌های تعیین میزان مقاومت روش‌های رگرسیون استوار در مقابل مشاهده‌های دورافتاده معیار نقطه فروریزش است که در واقع کوچک‌ترین کسری از تغییرات است که منجر می‌شود برآوردگر هر مقدار دلخواهی دور از مقدار اصلی برآورد را اتخاذ کند که این نسبت معمولاً به صورت مجانبی بیان می‌گردد. از جمله روش‌های رگرسیون استوار می‌توان به برآوردگر رگرسیونی کمترین قدر مطلق انحرافات^۸ (LAD) با

$$\hat{\beta}_{LAD} = \arg \min_{\beta} \sum_{i=1}^n |y_i - \mathbf{x}_i \beta|,$$

^۵ Leverage Point

^۶ Partial Leverage Plots

^۷ Robust

^۸ Least Absolute Deviations

^۹ Least Median of Squares

^{۱۰} Least Trimmed Squares

و استنباط‌های مربوط به ضرایب مدل رگرسیونی خطی را نامعتبر می‌سازد. یک راهکار برای کاهش این تورم انقباض و کشیدن برآورد به سمت صفر است که باعث کاهش واریانس برآورد گر نیز می‌شود.

همچنین در بعضی شرایط با یک مدل خطی تنک^{۱۴} مواجه هستیم. در یک مدل تنک تعدادی از ضرایب رگرسیونی برابر با صفر هستند. در اینگونه مسائل تشخیص و برآورد ضرایب رگرسیونی ناصفر از اهمیت خاصی برخوردار است و مسئله انتخاب متغیر مطرح می‌گردد. در مدل‌های دارای متغیرهای پیشگوی بسیار زیاد و یا مدل‌های تنک انقباض همراه با انتخاب متغیر می‌تواند به برآوردهای مناسبی منجر شود. منظم سازی^{۱۵} یا رگرسیون توانیده^{۱۶} راهکاری برای ایجاد انقباض در برآوردگرهای رگرسیونی است که با اضافه کردن یک جمله^{۱۷} توان به تابع هدف مدل رگرسیونی صورت می‌گیرد. برای مدل رگرسیونی با تابع هدف $Q(\mathbf{X}, \mathbf{y}, \beta)$ با اضافه کردن تابع توان $P(\beta)$ ، برآوردگر رگرسیونی توانیده^{۱۷} β به صورت

$$\hat{\beta}_{\text{Pen}} = \arg \min_{\beta} Q(\mathbf{X}, \mathbf{y}, \beta) + \lambda P(\beta).$$

به دست می‌آید. تابع توان $P(\beta)$ یک اندازه برای میزان بزرگی β و پارامتر $\lambda > 0$ برای تنظیم میزان تأثیر تابع توان در محدود سازی برآورد حاصل است. بنا به انتخاب تابع توان روش‌های رگرسیونی مختلفی ایجاد شده است که در این مقاله به سه روش پر کاربرد می‌پردازیم. این روش‌ها عبارت‌اند از روش رگرسیون ستیغی، روش کمترین قدر مطلق انقباض و عملگر انتخاب و روش تور مرتجع.

صفر درصد نمی‌رسد. در M -برآوردگرها در حقیقت تابع ρ وزن‌های در نظر گرفته برای مانده‌ها را کنترل می‌کند و بسته به نوع این تابع می‌توان نقطه‌های فروریزش متفاوتی را برای این برآوردگر به دست آورد. نقطه فروریزش مجانبی برآوردگر LMS، LTS و S -برآوردگرها (تحت شرایطی خاص) برابر با ۵۰ درصد است که بیشترین میزان نقطه فروریزش برای یک برآوردگر است [۱۶].

۳ هم‌خطی چندگانه، مدل‌های تنک و روش‌های رگرسیون توانیده

یکی دیگر از مشکلاتی که می‌تواند برآورد رگرسیونی کمترین توان‌های دوم را دچار اختلال کند وجود همبستگی میان متغیرهای پیشگو در مدل رگرسیونی است که به هم‌خطی چندگانه^{۱۱} معروف است. یک شاخص برای تشخیص هم‌خطی چندگانه، عامل تورم واریانس^{۱۲} (VIF) است. عامل تورم واریانس برای j امین ضریب رگرسیونی به صورت

$$VIF_j = \frac{1}{1 - R_j^2},$$

محاسبه می‌شود. که در آن R_j^2 ضریب تعیین^{۱۳} چندگانه رگرسیون $\mathbf{X}_j$ نسبت به دیگر متغیرهای پیشگوی مدل رگرسیونی است. عامل تورم واریانس بیش از ۱۰ نشان دهنده مشکلات جدی هم‌خطی چندگانه است.

هم‌خطی چندگانه باعث می‌شود دترمینان ماتریس $\mathbf{X}'\mathbf{X}$ بسیار به صفر نزدیک شود و در نتیجه $(\mathbf{X}'\mathbf{X})^{-1}$ مقادیر بسیار بزرگی را اختیار کند. از آنجا که $Cov(\hat{\beta}_{LS}) = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2$ ، این مسئله باعث تورم ماتریس واریانس کواریانس برآوردگرهای کمترین توان‌های دوم ضرایب شده

^{۱۱} Multicollinearity

^{۱۲} Variance Inflation Factor

^{۱۳} Coefficient of Determination

^{۱۴} Sparse

^{۱۵} Regularization

^{۱۶} Penalized R Plots

^{۱۷} Ridge Regression

۱.۳ روش رگرسیون ستیغی

رگرسیون ستیغی [۶ و ۷] یکی از متداول ترین تکنیک هایی است که در حالت هم خطی چندگانه مورد استفاده قرار می گیرد. در این روش تابع تاوان توان دوم نرم L_2 ی β یعنی $\beta'\beta$ است. برآوردگر ستیغی به صورت

$$\hat{\beta}_{\text{RIDGE}} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}, \quad (۶)$$

به دست می آید. که در آن عدد حقیقی نامنفی λ میزان بزرگی تابع تاوان را تنظیم می کند. برای این برآوردگر داریم

$$\hat{\beta}_{\text{RIDGE}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y}. \quad (۷)$$

به بیان دیگر برآوردگر ستیغی معادل است با مسئله مینیم سازی $\min_{\beta} \sum_{i=1}^n (y_i - \mathbf{x}_i \beta)^2$ تحت قید $\sum_{j=1}^p \beta_j^2 \leq M$ که در آن M یک کران ثابت است. در حقیقت یک رابطه یک به یک بین پارامتر λ در (۷) و M وجود دارد.

برآوردگر (۷) و (۷) تنها در یک جمله $\lambda \mathbf{I}$ که به ماتریس $\mathbf{X}'\mathbf{X}$ اضافه شده است اختلاف دارند. اضافه شدن مقدار مثبت λ به مؤلفه های قطری ماتریس $\mathbf{X}'\mathbf{X}$ باعث کاهش تورم معکوس آن می شود.

[۸] پیشنهاد کرده اند که پارامتر ستیغی λ باید به اندازه ای کوچک انتخاب شود که

$$\text{MSE}(\hat{\beta}_{\text{RIDGE}}) \leq \text{MSE}(\hat{\beta}_{\text{LS}})$$

و بر این اساس مقدار بهینه $\lambda = \frac{p\sigma^2}{\beta' \beta}$ را پیشنهاد داده اند، که پارامترهای β و σ^2 در آن رامی توان با $\hat{\beta}_{\text{LS}}$ و $\hat{\sigma}_{\text{LS}}^2$ جایگزین کرد.

۲.۳ روش کمترین قدر مطلق انقباض و عملگر انتخاب

یکی دیگر از روش های منظم سازی روش کمترین قدر مطلق انقباض و عملگر انتخاب ^{۱۸} (LASSO) است که توسط [۱۹] ارائه شده است.

در این روش تابع تاوان نرم L_1 بردار β است. برآوردگر LASSO به صورت

$$\hat{\beta}_{\text{LASSO}} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i \beta)^2 + n \lambda \sum_{j=1}^p |\beta_j| \right\}, \quad (۸)$$

به دست می آید. با توجه به مشتق ناپذیر بودن تابع تاوان $\sum_{j=1}^p |\beta_j|$ در نقطه صفر، مقادیر بزرگ تری از λ منجر به صفر شدن ضرایب رگرسیونی متغیرهای پیشگوی کم تأثیر در مدل می گردد و در نتیجه عمل برآورد پارامتر و انتخاب متغیر به صورت هم زمان انجام می گیرد.

به دلیل آنکه LASSO پارامترهای تنظیم یکسانی را برای همه ضرایب رگرسیونی به کار می برد، برآوردگرهای حاصل دارای اریبی قابل ملاحظه ای هستند. در راهکاری دیگر می توانیم با اختصاص دادن پارامترهای تنظیم متفاوت، تاوان کمتری برای ضرایب رگرسیونی متناظر با متغیرهای مهم در نظر گرفته و از اریبی برآوردگرهای حاصل بکاهیم. برآورد اصلاح شده ی LASSO که با نماد LASSO* نمایش داده می شود به صورت

$$\hat{\beta}_{\text{LASSO}^*} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i \beta)^2 + n \sum_{j=1}^p \lambda_j |\beta_j| \right\}.$$

به دست می آید. یک الگوریتم سریع به منظور محاسبه LASSO از طریق چهارچوب روش رگرسیونی کمترین زاویه ^{۱۹} (LARS) [۵] در دسترس است. رگرسیون کمترین زاویه مرتبط با روش کلاسیک انتخاب مدل رگرسیونی پیشروی مرحله به مرحله است. نکته قابل توجه این است که رابطه ای مستقیم بین پارامتر تنظیم روش LASSO و تعداد گام های روش LARS وجود دارد.

^{۱۸}Least Absolute Shrinkage and Selection Operator

^{۱۹} Least Angle Regression

که بیانگر این مطلب است که تور مرتجع ساده انگارانه به‌طور بالقوه در هر وضعیتی می‌تواند هر p متغیر را انتخاب کند. همچنین دلیل اصلی توانایی تور مرتجع در غلبه بر m از روش LASSO محذب بودن اکید تابع تاوان روش تور مرتجع نسبت به β ، به‌ازای $\lambda_2 > 0$ است [۲۰].

عملکرد برآوردگر تور مرتجع ساده انگارانه در دو مرحله صورت می‌گیرد، ابتدا برای هر λ_2 ثابت یک برآورد رگرسیونی ستیغی به دست آورده و سپس این برآورد را توسط تابع تاوان نرم L_1 منقبض می‌کند و هم‌زمان انتخاب متغیر را نیز انجام می‌دهد. واضح است که در این صورت یک انقباض مضاعف صورت گرفته است. با اصلاح این انقباض مضاعف، می‌توانیم عملکرد پیشگویی روش تور مرتجع ساده انگارانه را بهبود ببخشیم. این ایده منجر به مطرح شدن روش رگرسیونی تور مرتجع (اصلاح شده) به‌صورت زیر گردید [۲۰].

$$\hat{\beta}_{\text{EN}} = (1 + \lambda_2) \hat{\beta}_{\text{N-EN}}.$$

مسئله مهم دیگر در روش تور مرتجع انتخاب مقادیر مناسب برای پارامترهای تنظیم (λ_1, λ_2) است. همان‌طور که در مورد برآوردگر ستیغی بیان کردیم رابطه‌ای مستقیم بین پارامتر تنظیم λ_1 برآورد ستیغی با کران M در قید $\sum_{j=1}^p \beta_j^2 \leq M$ وجود دارد. بنا بر این پارامتر تنظیم λ_2 را می‌توان با $s = \frac{1}{M} \sum_{j=1}^p \beta_j^2$ جایگزین کرد. از طرف دیگر همان‌طور که گفتیم رابطه‌ای مستقیم بین پارامتر تنظیم روش LASSO و تعداد گام‌های (k) روش LARS وجود دارد. بنا بر این به جای (λ_1, λ_2) می‌توان از پارامترهای (k, s) استفاده کرد.

پارامتر تنظیم λ_1 (؟؟) به گونه‌ای برآورد می‌شود که یک معیار دقت برآورد رگرسیونی را حد اقل کند. معیارهای مختلفی برای سنجش دقت برآورد رگرسیونی وجود دارد که از جمله آنها می‌توان به معیار اطلاع آکائیکه^{۲۰} [۱]، معیار اطلاع بیزی^{۲۱} [۱۸] و معیار اعتبارسنجی متقابل^{۲۲} [۱۲] اشاره کرد.

۳.۳ روش تور مرتجع

با این که روش LASSO در بسیاری از تحلیل‌ها کارایی خوبی دارد، با این حال دارای محدودیت‌هایی است، که از جمله آنها می‌توان به موارد زیر اشاره کرد:

۱.م در حالت $p > n$ LASSO حد اکثر می‌تواند n متغیر انتخاب کند، که این مسئله یک ویژگی محدود کننده برای انتخاب متغیر است [۲۰].

۲.م اگر گروهی از متغیرها موجود باشد که همبستگی‌های دو به دوی آنها بسیار شدید باشد، در این صورت LASSO گرایش به انتخاب تنها یک متغیر از گروه دارد و متغیر انتخاب شده از این گروه لزوماً بهترین متغیر نیست [۲۰]. به‌منظور فائق آمدن بر این نقایص [۲۰] روش منظم سازی دیگری به نام تور مرتجع^{۲۳} را ارائه کرده‌اند. ابتدا حالت ساده انگارانه^{۲۴} این روش را بیان می‌کنیم.

برآوردگر ساده انگارانه تور مرتجع β به‌صورت

$$\hat{\beta}_{\text{N-EN}} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i \beta)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}.$$

به دست می‌آید. نحوه محاسبه برآوردگر تور مرتجع به این صورت است که با استفاده از داده‌های ساختگی می‌توان مسئله تور مرتجع ساده انگارانه را به یک مسئله LASSO تبدیل کرد [۲۰]. در حقیقت اندازه نمونه در مسئله ساختگی $n + p$ است،

^{۲۰} Akaike Information Criterion

^{۲۱} Bayesian Information Criterion

^{۲۲} Cross Validation

^{۲۳} Elastic Net

^{۲۴} Naive

و در نتیجه

$$\hat{\beta}_{\text{LTS-RIDGE}} = (\mathbf{X}'\mathbf{Z}^*\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{Z}^*\mathbf{y},$$

که در آن $\mathbf{Z}^* = \text{diag}(\mathbf{z}^*)$ و $\mathbf{z}^* = \arg \min_{\mathbf{z} \in \mathbf{E}_h} Q(\mathbf{z})$ که در آن $\mathbf{z} = (z_1, \dots, z_n)$ که $z_i \in \{0, 1\}$ ، $i = 1, \dots, n$

$$\mathbf{E}_h = \left\{ (z_1, \dots, z_n); z_i \in \{0, 1\}, \sum_{i=1}^n z_i = h \right\}$$

و

$$Q(\mathbf{z}) = \left(\mathbf{y} - \mathbf{X} (\mathbf{X}'\mathbf{Z}\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{Z}\mathbf{y} \right)' \mathbf{Z} \left(\mathbf{y} - \mathbf{X} (\mathbf{X}'\mathbf{Z}\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{Z}\mathbf{y} \right) + \lambda \mathbf{y}'\mathbf{Z}'\mathbf{X} (\mathbf{X}'\mathbf{Z}\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{Z}\mathbf{y}.$$

به منظور محاسبه برآورد LTS - RIDGE الگوریتم

FAST - LTS - RIDGE را به کار می‌بریم که به صورت

زیر است.

برای یک پارامتر تنظیم ثابت λ تابع هدف مجموع توان‌هایدوم تاوانیده با توان دوم نرم L_2 ی β روی زیرمجموعه $\mathbf{H} \subseteq \{1, \dots, n\}$ با $|\mathbf{H}| = h$ را به صورت

$$Q(\mathbf{H}, \beta) = \sum_{i \in \mathbf{H}} (y_i - \mathbf{x}_i\beta)^2 + \lambda \sum_{j=1}^p \beta_j^2.$$

در نظر بگیرید. برای انتخاب زیرمجموعه اولیه $(\mathbf{H}_0)$ به

این صورت عمل می‌کنیم که ابتدا زیرمجموعه‌ای دلخواه از

اندازه‌ی h انتخاب کرده و برای زیرمجموعه مذکور پارامتر λ

را به کمک معیار GCV برآورد می‌کنیم. قرار می‌دهیم

$$\mathbf{H}_1 = \left\{ (y_i, \mathbf{x}_i); e_{j:n}^{\dagger} \in \left\{ (e_{j:n}^{\dagger}) : j = 1, \dots, h \right\} \right\},$$

در مراحل بعدی الگوریتم برای برآورد پارامتر تنظیم λ از

روش برآورد پارامتر ارائه شده توسط [۸] که در بخش قبل به

آن اشاره کردیم استفاده می‌کنیم. در مرحله‌ی k ام بر اساسزیرمجموعه جاری $\mathbf{H}_k$ که $|\mathbf{H}_k| = h$ ، ساختن زیرمجموعهبعدی $\mathbf{H}_{k+1}$ از مشاهده‌های متناظر با h تا از کوچک‌ترین

توان دوم مانده‌های حاصل از به کار گرفتن برآورد گرستی

روی زیرمجموعه $\mathbf{H}_k$ صورت می‌پذیرد. همچنین می‌توان با

تکرار زیرمجموعه ابتدایی به یک مینیمم موضعی تا حد امکان

نزدیک به مینیمم مطلق دست یافت.

۴ برآورد گرستی کمترین توان‌های دوم پیراسته تاوانیده

روش‌های رگرسیونی تاوانیده معرفی شده در بخش قبل در برابر مشاهده‌های دورافتاده استوار نیستند. عدم استواری این روش‌ها ناشی از به کار گرفتن تابع هدف کمترین توان‌های دوم در ساختار تابع هدف آنها است. در این بخش با جایگزینی تابع هدف کمترین توان‌های دوم پیراسته (LTS) به جای تابع هدف کمترین توان‌های دوم نسخه‌هایی استوار از این روش‌های رگرسیون تاوانیده را به دست می‌آوریم. بدیهی است که این شیوه برآورد زمانی که به‌طور هم‌زمان با مشکل هم‌خطی چندگانه و یا تنک بودن مدل و مشاهده‌های دورافتاده مواجه باشیم کاربرد دارد.

۱.۴ روش رگرسیونی کمترین توان‌های دوم پیراسته ستیغی

روش کمترین توان‌های دوم پیراسته ستیغی (LTS - RIDGE) تلفیقی از روش رگرسیونی ستیغی و روش استوار کمترین توان‌های دوم پیراسته به‌منظور برطرف کردن مشکلات وجود هم‌خطی چندگانه و حضور مشاهده‌های دورافتاده به‌صورت هم‌زمان است، که با افزودن یک جمله تاوان توان دوم نرم L_2 ی β با پارامتر تنظیم λ به تابع هدف کمترین توان‌های دوم پیراسته به‌صورت

$$\hat{\beta}_{\text{LTS-RIDGE}} = \arg \min_{\beta} \left\{ \sum_{i=1}^h ((y - \mathbf{X}\beta)^{\dagger})_{i:n} + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

به دست می‌آید. که در آن

$$(y - \mathbf{X}\beta)_{1:n}^{\dagger} \leq \dots \leq (y - \mathbf{X}\beta)_{n:n}^{\dagger}$$

و h به ترتیب در رابطه‌های (؟؟) و (؟؟) داده شده‌اند.

به بیان دیگر می‌توان نوشت

$$\hat{\beta}_{\text{LTS-RIDGE}} = \arg \min_{\mathbf{z} \in \mathbf{E}_h} \arg \min_{\beta} \left\{ (\mathbf{y} - \mathbf{X}\beta)' \mathbf{Z} (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta' \beta \right\}.$$

نقطهٔ فروریزش این برآوردگر رگرسیونی LTS – LASSO استوار بالایی این روش است. $\frac{n-h+1}{n}$ است [۲]، که حاکی از استواری بالای این روش است.

همان‌طور که می‌دانیم روش رگرسیونی LTS از زیرمجموعه‌هایی از اندازه‌ی h استفاده می‌نماید، به بیان دیگر ممکن است اندازهٔ نمونهٔ h از تعداد متغیرهای p کمتر باشد، از این رو روش LTS روش مناسبی به‌منظور استفاده در کنار برآوردگر LASSO نیست، زیرا یکی از محدودیت‌های LASSO زمانی رخ می‌دهد که اندازهٔ نمونه از تعداد متغیرها کمتر باشد.

۳.۴ روش رگرسیونی کمترین توان‌های دوم پیراستهٔ تور مرتجع

به‌منظور فائق آمدن بر محدودیت‌های ناشی از روش LASSO که در زیربخش قبل بیان کردیم، [۱۳] روش رگرسیونی کمترین توان‌های دوم پیراستهٔ تور مرتجع را برای تعدیل مشکلات مشاهد‌های دورافتاده و هم‌خطی چندگانه مطرح کردند. برآوردگر رگرسیونی کمترین توان‌های دوم پیراستهٔ تور مرتجع ساده انگارانه $(LTS - N - EN)$ به‌صورت

$$\hat{\beta}_{LTS-N-EN} = \arg \min_{\beta} \left\{ \sum_{i=1}^h ((y - \mathbf{X}\beta)^{\top})_{i:n} + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}.$$

به دست می‌آید. که در آن

$$(y - \mathbf{X}\beta)_{1:n}^{\top} \leq \dots \leq (y - \mathbf{X}\beta)_{n:n}^{\top}$$

و h به ترتیب در رابطه‌های (؟؟) و (؟؟) داده شده‌اند. با توجه به این که روی هر زیرمجموعهٔ h تایی برآوردگر $LTS - N - EN$ تبدیل به یک برآوردگر $N - EN$ می‌شود، بنا بر این با استفاده از داده‌های ساختگی روی هر زیرمجموعهٔ h تایی برآوردگر $N - EN$ تبدیل به یک برآوردگر LASSO می‌شود. بنا بر این با الگوریتمی مشابه الگوریتم FAST – LTS – RIDGE و با در نظر گرفتن داده‌های ساختگی روی هر زیرمجموعهٔ h تایی می‌توان به برآوردگر $LTS - N - EN$ دست یافت.

نقطهٔ فروریزش این برآوردگر $\frac{n-h+1}{n}$ است که حاکی از استواری بالای این روش است.

[۱۱] روشی ساده انگارانه برای به دست آوردن یک برآورد کمترین توان‌های دوم پیراستهٔ ستیغی ارائه کرده‌اند که با توجه به رابطهٔ برآوردگر ستیغی و برآوردگر کمترین توان‌های دوم این برآوردگر به‌صورت

$$\hat{\beta}_{N-LTS-RIDGE} = (\mathbf{X}'\mathbf{X} + \lambda_{LTS}\mathbf{I})^{-1} \mathbf{X}'\mathbf{X}\hat{\beta}_{LTS},$$

به دست می‌آید. که در آن λ_{LTS} می‌تواند به‌صورت

$$\lambda_{LTS} = \frac{p\hat{\sigma}_{LTS}^2}{\hat{\beta}_{LTS}'\hat{\beta}_{LTS}}.$$

محاسبه شود.

۲.۴ روش رگرسیونی کمترین توان‌های دوم پیراستهٔ تنک

با وجود استواری بالای روش LTS واضح است که این روش برآوردهای مدل تنک را تولید نمی‌کند. یک نسخهٔ تنک از LTS با افزودن یک تابع تاوان نرم L_1 بردار β با پارامتر تنظیم λ به تابع هدف LTS حاصل می‌شود. برآوردگر کمترین توان‌های دوم پیراستهٔ تنک $(Sparse - LTS)$ به‌صورت

$$\beta_{Sparse-LTS} = \arg \min_{\beta} \left\{ \sum_{i=1}^h ((y - \mathbf{X}\beta)^{\top})_{i:n} + h\lambda \sum_{j=1}^p |\beta_j| \right\}.$$

به دست می‌آید. که در آن

$$(y - \mathbf{X}\beta)_{1:n}^{\top} \leq \dots \leq (y - \mathbf{X}\beta)_{n:n}^{\top}$$

و h به ترتیب در رابطه‌های (؟؟) و (؟؟) داده شده‌اند. برآوردگر LTS تنک می‌تواند به عنوان یک نسخهٔ پیراسته‌ی LASSO در نظر گرفته شود، از این رو نام دیگری که می‌توان برای این برآوردگر به کار برد LTS – LASSO است.

برای محاسبهٔ سریع برآوردگر LTS – LASSO می‌توان الگوریتمی مشابه الگوریتم FAST – LTS – RIDGE را به کار گرفت.

۵ رگرسیون استوار در حضور هم خطی

چند گانه به وسیله برآورگر کمترین

قدر مطلق انحرافات تاوانیده

اضافه کردن یک جمله تاوان به تابع هدف برآورگر LAD می تواند منجر به برآوردگرهای رگرسیونی استواری شود که در حضور هم خطی چند گانه و یا در مواجهه با مدل های تنک نیز دارای کارایی و عملکرد پیشگویی قابل اطمینانی هستند. در این بخش به بررسی برآوردگرهای LAD تاوانیده توسط تابع تاوان نرم L_1 و توان دوم نرم L_2 بردار ضرایب رگرسیونی و ترکیبی خطی از این دو نرم می پردازیم.

۱.۵ روش کمترین قدر مطلق انحرافات ستیغی

ایده اصلی روش کمترین قدر مطلق انحرافات ستیغی (LAD - RIDGE) افزودن یک تابع تاوان توان دوم نرم L_2 بردار ضرایب (RIDGE) به تابع هدف کمترین قدر مطلق انحرافات (LAD) است. این روش در مقایسه با روش کمترین قدر مطلق انحرافات می تواند عمل انقباض ضرایب رگرسیونی و برآورد پارامتر را به صورت هم زمان انجام دهد و در مقایسه با برآورگر ستیغی در مقابل مشاهده های دور افتاده عمودی استوار است. برآورگر LAD - RIDGE به صورت

$$\hat{\beta}_{\text{LAD-RIDGE}} = \arg \min_{\beta} \left\{ \sum_{i=1}^n |y_i - \mathbf{x}_i \beta| + \lambda \sum_{j=1}^p \beta_j^2 \right\}. \quad (9)$$

به دست می آید. به منظور محاسبه برآورگر LAD - RIDGE از تقریب خطی موضعی بر اساس بسط تیلور به صورت

$$\hat{\beta}_{\text{LAD-RIDGE}} \approx \arg \min_{\beta} \left\{ \sum_{i=1}^n |y_i - \mathbf{x}_i \beta| + n \sum_{j=1}^p \frac{2\lambda \beta_j^{(0)}}{n} |\beta_j| \right\}.$$

استفاده می کنیم. که در آن $\beta_j^{(0)}$ یک مقدار معلوم است. در نتیجه برآورگر LAD - RIDGE را می توان یک برآورگر $\text{LAD} - \text{LASSO}^*$ تقریبی در نظر گرفت، که در زیر بخش بعد به معرفی آن می پردازیم.

۲.۵ روش کمترین قدر مطلق انحرافات تنک

با افزودن یک تابع تاوان نرم L_1 بردار ضرایب رگرسیونی به تابع هدف LAD می توان از این روش در مواجهه با مدل های تنک بهره گرفت. این روش در مقایسه با LAD می تواند عمل انتخاب متغیر و برآورد پارامتر را به صورت هم زمان انجام دهد و در مقایسه با LASSO در مقابل مشاهده های دور افتاده عمودی استوار است.

همان طور که در بخش سوم اشاره کردیم، نسخه اصلاح شده LASSO ، که آن را با نماد LASSO^* نشان دادیم، دارای کارایی بالاتری نسبت به LASSO معمولی است. علیرغم پیچیدگی بیشتر LASSO^* نسبت به LASSO ، ترجیح می دهیم از ترکیب دو روش LAD و LASSO^* استفاده کنیم. برآورگر $\text{LAD} - \text{LASSO}^*$ به صورت

$$\hat{\beta}_{\text{LAD-LASSO}^*} = \arg \min_{\beta} \left\{ \sum_{i=1}^n |y_i - \mathbf{x}_i \beta| + n \sum_{j=1}^p \lambda_j |\beta_j| \right\}. \quad (10)$$

به دست می آید. از نظر محاسباتی مسئله یافتن برآورگر $\text{LAD} - \text{LASSO}^*$ با استفاده از داده های ساختگی به یک مسئله یافتن یک برآورگر LAD کاهش می یابد. به منظور محاسبه برآورگر LAD می توان از الگوریتم BR [۳] استفاده کرد.

به دلیل وجود p پارامتر تنظیم در مدل (۹) و برآورد هم زمان این پارامترها، استفاده از روش های CV، GCV و BIC برای برآورد این پارامترها مناسب نیست. بر اساس ایده ای از [۱۹] برآورگر $\text{LAD} - \text{LASSO}^*$ را می توان به عنوان یک برآورگر بیزی در نظر گرفت، به طوری که هر ضریب رگرسیونی β_j دارای توزیع پیشین نمایی دو گانه با پارامتر مکانی صفر و پارامتر مقیاسی $n\lambda_j$ است و بر این اساس مقدار بهینه $\hat{\beta}_j = \frac{1}{n\lambda_j}$ را ارائه شده است، که در آن β_j را می توان با استفاده از برآورگر LAD معمولی برآورد کرد.

۳.۵ روش کمترین قدر مطلق انحرافات تور مرتجع

ایده اصلی روش کمترین قدر مطلق انحرافات تور مرتجع (LAD - EN) افزودن یک تابع تاوان از نوع تور مرتجع به تابع هدف کمترین قدر مطلق انحرافات است. این روش در مقایسه با روش کمترین قدر مطلق انحرافات می‌تواند عمل تنک سازی و انقباض ضرایب رگرسیونی و برآورد پارامتر را به صورت هم‌زمان انجام دهد و در مقایسه با برآوردگر تور مرتجع در مقابل مشاهده‌های دورافتاده عمودی استوار است. همچنین این روش معایب روش LAD - LASSO در انتخاب متغیر را ندارد. برآوردگر کمترین قدر مطلق انحرافات تور مرتجع ساده انگارانه به صورت

$$\hat{\beta}_{\text{LAD-N-EN}} = \arg \min_{\beta} \left\{ \sum_{i=1}^n |y_i - \mathbf{x}_i \beta| + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}.$$

به دست می‌آید. به منظور سهولت محاسبه برآوردگر LAD - N - EN، از تقریب

$$\hat{\beta}_{\text{LAD-N-EN}} \approx \arg \min_{\beta} \left\{ \sum_{i=1}^n |y_i - \mathbf{x}_i \beta| + n \sum_{j=1}^p \left(\frac{2\lambda_2 \beta_j^{(2)} + \lambda_1}{n} \right) |\beta_j| \right\}.$$

استفاده می‌کنیم. در نتیجه برآوردگر LAD - N - EN به یک برآوردگر LAD - LASSO* تقریبی تبدیل می‌شود.

۶ مطالعه عددی، شبیه‌سازی و بررسی داده‌های واقعی

به منظور مقایسه عملکرد برآوردگرهای مختلف مطرح شده در بخش‌های چهارم و پنجم ابتدا سه مطالعه شبیه‌سازی مجزا را انجام می‌دهیم. سپس به تحلیل یک مجموعه داده واقعی چربی

بدن [۱۰] با استفاده از روش‌های رگرسیون استوار تاوانیده‌ی LTS - LASSO و LTS - RIDGE می‌پردازیم.

۱.۶ شبیه‌سازی

در این بخش به مطالعه سه شبیه‌سازی مجزا برای مقایسه عملکرد برآوردگرهای مختلف مطرح شده در بخش‌های چهارم و پنجم می‌پردازیم. این شبیه‌سازی‌ها در نرم‌افزار R انجام شده است. هدف از شبیه‌سازی اول مقایسه عملکرد روش‌های LASSO و EN برای تنک سازی برآوردگرهای به دست آمده از یک طرف و همچنین مقایسه روش‌های LAD و LTS برای استوار سازی برآوردهای حاصل در برابر مشاهده‌های دورافتاده از طرف دیگر است. به این منظور چهار ترکیب مختلف از این روش‌ها را به صورت LAD - LASSO*، LTS - LASSO، LAD - EN و LTS - EN در نظر می‌گیریم. علاوه بر محاسبه و مقایسه عملکرد این چهار برآوردگر، دو روش LTS و LAD را نیز به منظور بررسی میزان تأثیرگذاری توابع تاوان در بهبود برآوردگرهای حاصل و دو روش LASSO و EN را نیز به منظور بررسی روش‌های استوار سازی در دستور کار قرار می‌دهیم. همچنین عملکرد برآورد کمترین توان‌های دوم (LS) را نیز در کنار برآوردگرهای بالا بررسی می‌کنیم. ملاک‌های بررسی دقت عملکرد برآوردگرها در این مطالعه شبیه‌سازی، ملاک ریشه دوم میانگین توان‌های دوم خطای پیشگویی^{۲۵} و ملاک‌های نرخ تنک نکردن نادرست و نرخ تنک سازی نادرست هستند که به ترتیب، نرخ مثبت نادرست^{۲۶} و نرخ منفی نادرست^{۲۷} نامیده می‌شوند و به صورت زیر تعریف می‌گردند.

ریشه دوم میانگین توان‌های دوم خطای پیشگویی (RMSPE): به منظور محاسبه این ملاک، ابتدا n مشاهده اضافی (y_i^*, x_i^*) ، $i = 1, \dots, n$ ، از مدل مورد نظر، بدون مشاهده‌های دورافتاده، به عنوان داده آزمایشی تولید می‌کنیم.

^{۲۵} Root Mean Squared Prediction Error

^{۲۶} False Positive Rate

^{۲۷} False Negative Rate

۱.۱.۶ شبیه‌سازی اول

در هر بار تکرار این شبیه‌سازی تعداد $n = ۱۰۰$ مشاهده متغیرهای پیشگو از یک توزیع نرمال p متغیره‌ی $N_p(0, I_{p \times p})$ تولید شده است و بردار ضرایب رگرسیونی به صورت $\beta = (10, 0, 0, 15, 0)$ داده شده است. در نتیجه در این شبیه‌سازی با یک مدل تنک مواجه هستیم. متغیرهای پاسخ طبق مدل $y_i = x_i\beta + \varepsilon_i$ تولید شده است، که در آن خطاهای ε_i از توزیع نرمال $N(0, 0.5)$ تولید شده‌اند.

به منظور تولید مشاهده‌های دورافتاده سه طرح پیشنهادی زیر را در نظر می‌گیریم:

(۱). عدم ایجاد مشاهده‌های دورافتاده.

(۲). ایجاد مشاهده‌های دورافتاده عمودی: ۱۰٪ از خطاهای مدل به جای توزیع $N(0, 0.5)$ از توزیع $N(40, 0.5)$ تولید شده‌اند.

(۳). ایجاد مشاهده‌های دورافتاده نافذ: مشابه (۲)، با این تفاوت که ۱۰٪ از متغیرهای پیشگوی متناظر با طرح (۲) نیز از توزیع‌های مستقل $N(2, 1)$ تولید شده‌اند.

جدول‌های ۱، ۲ و ۳ میانگین RMSPE، FPR و FNR ها را برای برآوردهای مختلف محاسبه شده در این شبیه‌سازی و به ترتیب به‌ازای سه طرح مختلف «بدون دورافتاده» (طرح (۱))، «دورافتاده عمودی» (طرح (۲)) و «دورافتاده نافذ» (طرح (۳)) نشان می‌دهد.

همان‌طور که از جدول ۱ قابل مشاهده است، در صورت عدم وجود مشاهده‌های دورافتاده، با توجه به این که با یک مدل تنک مواجه هستیم، برآوردهای EN و LASSO عملکرد پیشگویی خوبی را از دیدگاه RMSPE و FPR از خود نشان داده‌اند. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی برآوردهای LS، LAD و LTS با توجه به کاهش میزان RMSPE و FPR برآوردهای توانانیده بر اساس این سه برآوردها مشهود است.

حال ملاک RMSPE را به صورت

$$\text{RMSPE}(\hat{\beta}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^* - x_i^* \hat{\beta})^2},$$

محاسبه می‌کنیم. نرخ مثبت نادرست (FPR): این ملاک در واقع نسبت ضرابی است که صفر هستند ولی روش تنک سازی موفق به صفر کردن مقدار برآورد آنها نشده است و به صورت

$$\text{FPR}(\hat{\beta}) = \frac{|\{j \in \{1, \dots, p\} : \hat{\beta}_j \neq 0, \beta_j = 0\}|}{|\{j \in \{1, \dots, p\} : \beta_j = 0\}|}.$$

به دست می‌آید. نرخ منفی نادرست (FNR): این ملاک در واقع نسبت ضرابی است که به اشتباه برآورد آنها صفر شده است و به صورت

$$\text{FNR}(\hat{\beta}) = \frac{|\{j \in \{1, \dots, p\} : \hat{\beta}_j = 0, \beta_j \neq 0\}|}{|\{j \in \{1, \dots, p\} : \beta_j \neq 0\}|}.$$

محاسبه می‌شود. در نهایت میانگین مقادیر RMSPE، FPR و FNR روی تعداد تکرارهای شبیه‌سازی محاسبه و گزارش می‌شوند.

هدف از شبیه‌سازی دوم مقایسه عملکرد روش‌های برآورد ستیغی استوار مختلف است. به این منظور سه روش LTS – RIDGE، LAD – RIDGE، ساده انگارانه و LTS – RIDGE دقیق را مورد مطالعه و مقایسه قرار می‌دهیم. علاوه بر محاسبه و مقایسه عملکرد این سه برآوردها، دو روش LTS و LAD را نیز به منظور بررسی میزان تأثیرگذاری اریب سازی ستیغی در بهبود برآوردهای حاصل و روش RIDGE را نیز به منظور بررسی روش‌های استوار سازی در دستور کار قرار می‌دهیم. همچنین عملکرد برآورد کمترین توان‌های دوم (LS) را نیز در کنار برآوردهای بالا بررسی می‌کنیم. در این شبیه‌سازی تنها ملاک بررسی دقت عملکرد برآوردها، ریشه دوم میانگین توان‌های دوم خطای پیشگویی است. هدف از شبیه‌سازی سوم مقایسه همه روش‌های مطرح شده در دو شبیه‌سازی قبل در مواجهه با یک مدل تنک و حضور هم‌خطی چندگانه است. ملاک‌های بررسی دقت عملکرد برآوردها در این مطالعه شبیه‌سازی نیز مشابه شبیه‌سازی است.

جدول ۱. میانگین RMSPE، FPR و FNRهای شبیه‌سازی اول برای طرح (۱)

	LS	LTS	LAD	LASSO	EN	LTS – LASSO	LAD – LASSO*	LTS – EN	LAD – EN
RMSPE	۰/۵۲۰	۰/۵۷۱	۰/۵۳۴	۰/۵۱۱	۰/۵۱۰	۰/۵۴۱	۰/۵۱۴	۰/۵۶۵	۰/۵۳۳
FPR	۱	۱	۱	۰/۱۳۳۳	۰/۱۸۳	۰/۵۸۳	۰/۲۸۳	۰/۹۱۷	۰/۹۳۳
FNR	۰	۰	۰	۰	۰	۰	۰	۰	۰

جدول ۲. میانگین RMSPE، FPR و FNRهای شبیه‌سازی اول برای طرح (۲)

	LS	LTS	LAD	LASSO	EN	LTS – LASSO	LAD – LASSO*	LTS – EN	LAD – EN
RMSPE	۵/۱۲۵	۰/۶۰۷	۰/۵۵۷	۳/۹۲۳۴	۸/۹۴۵	۰/۵۴۶	۰/۵۳۲	۰/۵۶۷	۰/۵۵۰
FPR	۱	۱	۱	۰/۰۶۷	۰/۰۶۷	۰/۵۱۷	۰/۳۸۳	۰/۷۱۷	۰/۹۶۷
FNR	۰	۰	۰	۰	۰	۰	۰	۰	۰

جدول ۳. میانگین RMSPE، FPR و FNRهای شبیه‌سازی اول برای طرح (۳)

	LS	LTS	LAD	LASSO	EN	LTS – LASSO	LAD – LASSO*	LTS – EN	LAD – EN
RMSPE	۷/۲۵۴	۰/۶۰۴	۰/۷۲۷	۳/۸۰۹	۹/۳۶۶	۰/۵۵۶	۰/۶۳۹۶	۲/۶۷۰	۰/۷۰۳
FPR	۱	۱	۱	۰/۶۶۷	۰/۹۹۹	۰/۶	۰/۸	۰/۷۳۳	۰/۹۹۹۹
FNR	۰	۰	۰	۰	۰	۰	۰	۰	۰

جدول ۴. میانگین RMSPEها برای شبیه‌سازی دوم

	LS	LTS	LAD	RIDGE	LTS – RIDGE1	LTS – RIDGE2	LAD – RIDGE
بدون دورافتاده	۰/۵۱۷	۰/۵۵۱	۰/۵۲۳	۰/۵۱۷	۰/۵۵۰	۰/۵۴۹	۰/۵۲۱
دورافتاده عمودی	۴/۷۳۶	۰/۵۵۵	۰/۵۵۲	۴/۳۹۰	۰/۷۸۱	۰/۵۴۶	۰/۶۳۴
دورافتاده نافذ	۰/۶۸۹۹	۰/۵۶۷	۰/۸۴۹	۵/۶۴۱	۰/۶۰۸	۰/۵۶۲	۰/۷۷۵

با توجه به جدول ۲ در صورت وجود مشاهده‌های دورافتاده عمودی در مجموعه داده مورد بررسی، همان‌طور که انتظار داشتیم برآوردهای $LAD - LASSO^*$ و $LTS - LASSO$ دارای عملکرد پیشگویی خوبی از دیدگاه RMSPE هستند. مقدار RMSPE بالای دو روش $LASSO$ و EN از عدم استواری این دو روش در مقابل مشاهده‌های دورافتاده عمودی ناشی شده است، با این حال این دو روش دارای FPR کمتری نسبت به دیگر روش‌های مورد بررسی هستند. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی برآوردهای LS ،

LAD و LTS با توجه به کاهش میزان RMSPE و FPR برآوردهای توانیده بر اساس این سه برآوردها مشهود است.

با توجه به جدول ۳ در صورت وجود هم‌زمان مشاهده‌های دورافتاده عمودی و نقاط نافذ، همان‌طور که انتظار داشتیم، برآوردهای $LTS - LASSO$ دارای بهترین عملکرد پیشگویی از دیدگاه RMSPE و FPR در مقایسه با دیگر برآوردهای توانیده در این شبیه‌سازی است. به دلیل عدم استواری برآوردهای LS ، دو برآوردهای $LASSO$ و EN عملکرد

دارای بهترین عملکرد پیشگویی از دیدگاه RMSPE هستند. RMSPE بالایی برای دو برآوردگر LS و RIDGE از عدم استواری این دو برآوردگر در مقابل مشاهده‌های دورافتاده عمودی ناشی شده است. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی برآوردگرهای LS، LAD و LTS با توجه به کاهش میزان RMSPE برآوردگرهای توانانیده بر اساس این سه برآوردگر مشهود است. در صورت وجود هم‌زمان مشاهده‌های دورافتاده عمودی و نقاط نافذ، همان‌طور که قابل مشاهده است، برآوردگر $LTS - RIDGE2$ دارای بهترین عملکرد پیشگویی از دیدگاه RMSPE است. RMSPE بالایی برای برآوردگرهای LS و RIDGE ناشی از عدم استواری این برآوردگرها در مقابل مشاهده‌های دورافتاده عمودی و نقاط نافذ است. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی برآوردگرهای LS، LAD و LTS، با توجه به کاهش میزان RMSPE برآوردگرهای توانانیده بر اساس این سه برآوردگر مشهود است.

۳.۱.۶ شبیه‌سازی سوم

در هر بار تکرار این شبیه‌سازی نیز تعداد $n = 100$ مشاهده متغیرهای پیشگو از یک توزیع نرمال مشابه شبیه‌سازی دوم تولید شده است و بردار ضرایب رگرسیونی نیز مشابه شبیه‌سازی اول به صورت $\beta = (10, 0, 0, 15, 0)$ داده شده است. در نتیجه در این شبیه‌سازی با یک مدل تنک همراه با هم‌خطی چندگانه مواجه هستیم. متغیرهای پاسخ نیز مشابه دو شبیه‌سازی اول و دوم طبق مدل $y_i = x_i\beta + \varepsilon_i$ تولید شده است، که در آن خطاهای ε_i از توزیع نرمال $N(0, 0.5)$ تولید شده‌اند. سه طرح پیشنهادی برای تولید مشاهده‌های دورافتاده در شبیه‌سازی اول و دوم را در این شبیه‌سازی نیز به کار برده‌ایم. جدول‌های ۵ و ۶ میانگین RMSPE، FPR و FNR ها را برای برآوردگرهای مختلف محاسبه شده در این شبیه‌سازی و به ترتیب به‌ازای دو طرح مختلف «دورافتاده عمودی» (طرح (۲)) و «دورافتاده نافذ» (طرح (۳)) نشان می‌دهد.

پیشگویی خوبی را از خود نشان نداده‌اند. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی برآوردگرهای LS، LAD و LTS با توجه به کاهش میزان FPR برآوردگرهای توانانیده بر اساس این سه برآوردگر مشهود است. همان‌طور که از ستون مربوط به مقادیر FNR در هر سه جدول فوق مشخص است، هیچ یک از روش‌های مورد بررسی در این شبیه‌سازی، منجر به صفر شدن نادرست برآورد ضرایب رگرسیونی در این شبیه‌سازی نشده‌اند.

۲.۱.۶ شبیه‌سازی دوم

شبیه‌سازی دوم مشابه شبیه‌سازی اول است، با این تفاوت که ماتریس واریانس کواریانس متغیرهای پیشگو به صورت $\Sigma = (\sigma_{ij})_{1 \leq i, j \leq p}$ با $\sigma_{ij} = 0.99^{|i-j|}$ و بردار ضرایب رگرسیونی به صورت $\beta = (10, 5, 8, 15, 4)$ داده شده است. در نتیجه در شبیه‌سازی دوم مدل دارای هم‌خطی چندگانه است. سه طرح پیشنهادی برای تولید مشاهده‌های دورافتاده در شبیه‌سازی اول را در این شبیه‌سازی نیز به کار برده‌ایم. مقادیر RMSPE برای برآوردگرهای مختلف و طرح‌های مختلف مشاهده‌های دورافتاده در جدول ۴ داده شده است.

همان‌طور که از جدول ۴ می‌توان مشاهده کرد، در صورت عدم وجود مشاهده‌های دورافتاده، با توجه به وجود هم‌خطی چندگانه در مدل، همان‌طور که انتظار داشتیم برآوردگر RIDGE دارای کمترین مقدار RMSPE است و در رتبه بعدی برآوردگر LS قرار دارد. به‌طور کلی با توجه به این که برآورد پارامتر تنظیم برآوردگر RIDGE به نحوی صورت می‌پذیرد که $MSE(\hat{\beta}_{RIDGE}) \leq MSE(\hat{\beta}_{LS})$ ، واضح است که در صورت وجود هم‌خطی چندگانه در مدل، روش RIDGE بر روش LS ارجحیت دارد. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی برآوردگرهای LS، LAD و LTS با توجه به کاهش میزان RMSPE برآوردگرهای توانانیده بر اساس این سه برآوردگر مشهود است. در صورت وجود مشاهده‌های دورافتاده عمودی همان‌طور قابل مشاهده است، دو برآوردگر $LAD - RIDGE$ و $LTS - RIDGE2$

۲.۶ بررسی داده واقعی

این مثال مربوط به داده‌های واقعی چربی بدن است که حاصل از مطالعه‌ای به منظور برآورد چگالی بدن ۲۵۲ مرد بر اساس اندازه‌گیری اندازه‌های مختلف اندام آنها صورت گرفته است. این مجموعه داده در بسته نرم‌افزاری mfp در نرم‌افزار R موجود است. متغیر پاسخ در این مطالعه میزان چگالی بدن است. همچنین متغیرهای پیشگوی در نظر گرفته شده شامل وزن، قد، اندازه‌ی دور گردن، اندازه‌ی دور سینه، اندازه‌ی دور شکم، اندازه‌ی دور مفصل ران، اندازه‌ی دور ران، اندازه‌ی دور زانو، اندازه‌ی قوزک پا، اندازه‌ی دو سر بازو، اندازه‌ی دور بازو و اندازه‌ی دور مچ هستند. مدل رگرسیونی خطی به صورت

$$\begin{aligned} \text{density} = & \beta_0 + \beta_1 \text{Weight} + \beta_2 \text{Height} + \beta_3 \text{Neck} \\ & + \beta_4 \text{Chest} + \beta_5 \text{Abdomen} + \beta_6 \text{Hip} \\ & + \beta_7 \text{Thigh} + \beta_8 \text{Knee} + \beta_9 \text{Ankle} + \beta_{10} \text{Biceps} \\ & + \beta_{11} \text{Forearm} + \beta_{12} \text{Wrist} + \varepsilon. \end{aligned}$$

برای این بررسی در نظر گرفته شده است. برای پی بردن به میزان وابستگی خطی میان این متغیرها مقادیر عوامل تورم واریانس (VIF) را برای این متغیرها محاسبه کرده‌ایم که به صورت زیر است.

با توجه به مقادیر VIF بالا به خصوص برای متغیرهای Weight، Chest، Abdomen و Hip می‌توان مشاهده کرد که هم‌خطی چندگانه شدیدی بین این متغیرها وجود دارد. از طرف دیگر نمودارهای جزئی نقاط نافذ که در شکل ۱ آورده شده‌اند، وجود برخی مشاهده‌های دورافتاده را در این مجموعه داده نشان می‌دهند.

برآورد ضرایب رگرسیونی به روش‌های LTS – LASSO و LTS – RIDGE دقیق به ترتیب به شرح زیر است

همان‌طور که از جدول ۵ قابل مشاهده است، در صورت وجود مشاهده‌های دورافتاده عمودی دو برآوردگر $LAD - LASSO$ و $LTS - LASSO$ دارای عملکرد پیشگویی خوبی از دیدگاه RMSPE هستند. RMSPE بالای دو روش RIDGE و EN از عدم استواری این دو روش در مقابل مشاهده‌های دورافتاده عمودی و نیز عدم قابلیت برآوردگر RIDGE در تنک سازی است. RMSPE بالای روش LS نیز ناشی از عدم استواری این روش و عدم توانایی در تنک سازی و غلبه بر هم‌خطی چندگانه است. همچنین روش LASSO با وجود داشتن قابلیت تنک سازی ناتوان از غلبه بر مشاهده‌های دورافتاده و مشکل هم‌خطی چندگانه موجود در مدل مورد بررسی است و در نتیجه RMSPE بالایی را اختیار کرده است. همان‌طور که در جدول ۵ قابل مشاهده است روش‌هایی که قابلیت تنک سازی را دارا هستند نسبت به روش‌هایی که قابلیت غلبه بر هم‌خطی چندگانه را دارند عملکرد بهتری را از دیدگاه FPR از خود نشان داده‌اند.

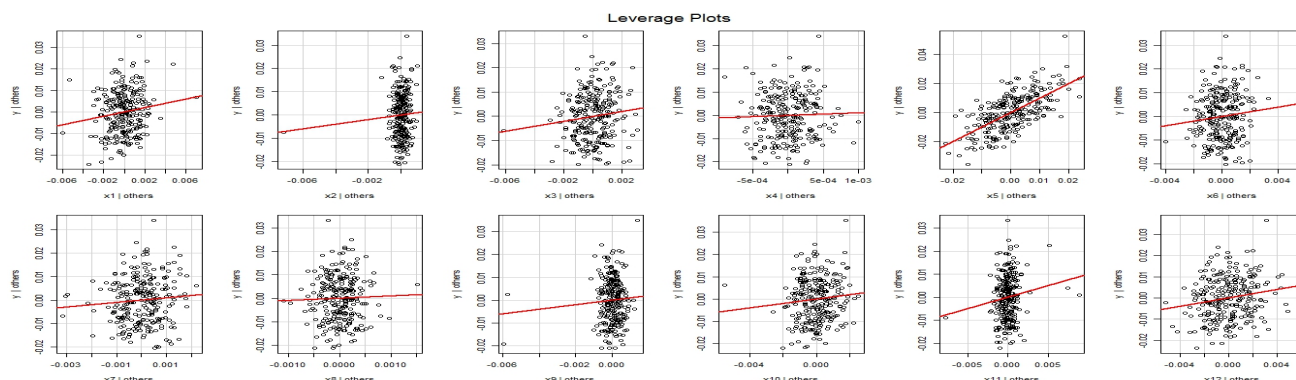
همان‌طور که در جدول ۶ قابل مشاهده است دو روش LTS – LASSO و LTS دارای عملکرد پیشگویی خوبی از دیدگاه RMSPE هستند که شاید RMSPE کم LTS ناشی از استواری بالای این روش در مقابل دورافتاده‌های ایجاد شده در طرح (۳) است. با این حال LTS به دلیل عدم قابلیت تنک سازی عملکرد خوبی را از دیدگاه FPR از خود نشان نداده است و در تشخیص درست ضرایب رگرسیونی صفر دچار خطا شده است و در مقابل برآوردگر LTS – LASSO به علت دارا بودن این قابلیت عملکرد خوبی را از دیدگاه FPR و همچنین FNR از خود نشان داده است.

جدول ۵. میانگین RMSPE، FPR و FNR های شبیه سازی سوم برای طرح (۲)

	RMSPE	FPR	FNR
LS	۵/۶۸۸۶۴	۱	۰
LTS	۰/۵۶۱۵۹	۱	۰
LAD	۰/۵۵۳۳۶	۱	۰
LASSO	۴/۵۴۱	۰/۳۵	۰/۲۲۵
EN	۸/۵۷۶۶۷	۰/۵۸۳۳۳	۰/۱۵
LTS – LASSO	۰/۵۵۱۹۱	۰/۵۶۶۶۷	۰
LAD – LASSO*	۰/۵۲۳۹۸	۰/۷۵	۰
LTS – EN	۰/۵۵۳۹۳	۰/۶۶۶۶۷	۰
LAD – EN	۰/۷۳۰۰۳	۰/۹۸۳۳۳	۰
RIDGE	۴/۸۳۶۶۵	۱	۰
LTS – RIDGE – 1	۱/۳۲۰۷۳	۱	۰
LTS – RIDGE – 2	۰/۵۶۳۱۳	۱	۰
LAD – RIDGE	۰/۷۳۲۹۷	۱	۰

جدول ۶. میانگین RMSPE، FPR و FNR های شبیه سازی سوم برای طرح (۳)

	RMSPE	FPR	FNR
LS	۵/۱۲۱۹۳	۱	۰
LTS	۱/۰۰۵۵۸	۱	۰
LAD	۲/۷۰۶۹۶	۱	۰
LASSO	۵/۰۳۷۱۱	۰/۵۵	۰/۱۵
EN	۷/۹۳۲۹۹	۰/۸۵	۰/۰۵
LTS – LASSO	۰/۷۴۲۸۱	۰/۵۱۶۶۷	۰
LAD – LASSO*	۲/۵۸۲۲۹	۱	۰/۰۲۵
LTS – EN	۱/۳۷۳۹۶	۰/۸۱۶۶۷	۰/۰۷۵
LAD – EN	۱/۸۲۸۶	۰/۹۵	۰/۰۲۵
RIDGE	۴/۹۳۳۵۱	۱	۰
LTS – RIDGE – 1	۱/۵۱۳۴۲	۱	۰
LTS – RIDGE – 2	۱/۰۱۷۶۲	۱	۰
LAD – RIDGE	۱/۸۳۱۲۵	۰/۹۵	۰/۰۲۵



شکل ۱. نمودارهای جزئی نقاط نافذ، مربوط به داده‌های واقعی چربی بدن

	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
VIF	۳۲/۰۷	۱/۶۵	۴/۲۹	۹/۴۵	۹/۳۸	۱۴/۷۰	۶/۷۴	۴/۳۲	۱/۸۹	۳/۶۰	۲/۱۳	۲/۹۹

متغیرها	Intercept	Weight	Height	Neck	Chest	Abdomen	Hip
$\hat{\beta}_{\text{LTS-LASSO}}$	۱/۱۰۰۱۳	-۰/۰۰۰۰۳	۰/۰۰۰۲۸	۰/۰۰۱۵۳	-۰/۰۰۰۵۶	-۰/۰۰۱۴۹	۰/۰۰۰۲۹

متغیرها	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
$\hat{\beta}_{\text{LTS-LASSO}}$	-۰/۰۰۰۳۴	۰/۰۰۰۶۷	۰/۰۰۰۵۴	۰/۰۰۱۲۹	-۰/۰۰۰۳۹	۰

متغیرها	Intercept	Weight	Height	Neck	Chest	Abdomen	Hip
$\hat{\beta}_{\text{LTS-RIDGE}}$	۱/۱۷۱۰۲	۰/۰۰۰۱۶	۰/۰۰۰۳۶	۰/۰۰۰۷۲	-۰/۰۰۰۶۶	-۰/۰۰۱۶۲	۰/۰۰۰۰۲

متغیرها	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
$\hat{\beta}_{\text{LTS-RIDGE}}$	۰/۰۰۰۶۶-	۰/۰۰۱۸۰	۰/۰۰۰۳۵	۰/۰۰۱۲۶	۰/۰۰۰۷۸-	۰/۰۰۲۲۱-

۷ بحث و نتیجه گیری

می‌توان توابع تاوان دیگری را در ساختار تابع هدف این برآوردگرها به کار گرفت. توابع تاوان غیر مقعر و از جمله مهم‌ترین و بهترین آنها تابع تاوان SCAD گزینه مناسبی به این منظور است که در کنار یک روش رگرسیونی استوار با نقطه فرویزش بالا برآوردگرهای رگرسیونی استوار تاوانیده مناسبی را در اختیار ما قرار می‌دهد. بدیهی است که می‌توان کارهای انجام شده در این مقاله را برای تابع تاوان SCAD نیز مورد بررسی قرار داد و به نتایج خوب و قابل قبولی دست یافت.

در این مقاله پس از بررسی نقایص برآوردگر رگرسیونی کلاسیک کمترین توان‌های دوم در برابر مشاهد‌های دورافتاده، هم‌خطی چندگانه و مواجهه با مدل‌های تنک، به معرفی روش‌های رگرسیونی استوار تاوانیده بر اساس سه نوع متداول تابع تاوان به عنوان راهکاری برای فائق آمدن بر این مشکلات پرداختیم و به نتایج قابل قبولی نیز دست یافتیم. با این حال به منظور بهبود هر چه بیشتر عملکرد این برآوردگرها و کاهش میزان اریبی و واریانس و افزایش کارایی آنها

مراجع

- [1] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *In 2nd International Symposium on Information Theory*. Akademia Kiado, 267–281.
- [2] Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, **7**(1), 226-248.
- [3] Barrodale, I., and Roberts, F. D. K. (1977). Algorithms for restricted least absolute value estimation. *Communications in Statistics-Simulation and Computation*, **6**(4), 353-363.
- [4] Edgeworth, F. Y. (1887). On observations relating to several quantities, *Hermathena*, **6**, 279-285.
- [5] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407–499.
- [6] Hoerl, A. E. and Kennard, R. W. (1970a). Ridge Regression: Iterative Estimation of the Biasing Parameter. *Communications in Statistics - Theory and Methods*, **5**, 77-88.
- [7] Hoerl, A. E. and Kennard, R. W. (1970b). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12**, 69-82.
- [8] Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: Some simulation, *Communications in Statistics - Theory and Methods*, **4**, 105-123.
- [9] Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Applied Statistics*, **1**, 799-821.
- [10] Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, **4**(1), 265-266.
- [11] Kan, B. and Alpu, Ö. and Yazıcı, B. (2013). Robust ridge and robust Liu estimator for regression based on the LTS estimator, *Journal of Applied Statistics*, **40**(3), 644-655.
- [12] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In Ijcai.*, **14**(2), 1137-1145.
- [13] Park, H., Sakaori, F. and Konishi, S. (2014). Robust sparse regression and tuning parameter selection via the efficient bootstrap information criteria. *Journal of Statistical Computation and Simulation*, **84**(7), 1596-1607.
- [14] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871-880.

- [15] Rousseeuw, P. J. (1984). Multivariate Estimation With High Breakdown Point. *Mathematical Statistics and Applications*, **8**, 283-297.
- [16] Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and outlier detection*. Wiley, New York.
- [17] Rousseeuw, P. J. and Yohai, V. (1984). Robust regression by means of S-estimators. *In Robust and Nonlinear Time Series Analysis*. Springer US, 256-272.
- [18] Schwarz, Gideon E. (1978), Estimating the dimension of a model. *Annals of Statistics*, **6(2)**, 461–464.
- [19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, **58**, 267–288.
- [20] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, **67**, 301–320.