براورد استوار نسبت به مشاهده‌های دورافتاده در رگرسیون خطی در حضور هم‌خطی 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 مشاهده‌های دورافتاده است. به روش‌هایی که تحت تأثیر مشاهده‌های دورافتاده قرار نمی‌گیرند، روش‌های آماری استوار گفته می‌شود. علاوه بر وجود مشاهده‌های دورافتاده، وجود وابستگی خطی میان متغیرهای پیشگو، که از آن با عنوان هم‌خطی چندگانه یاد می‌شود و نیز تعداد زیاد متغیرها در مقابل اندازهٔ کم نمونه، به خصوص در مدل‌های تُنک 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 مشاهده‌های دورافتاده، هم‌خطی چندگانه و مدل‌های تنک را بررسی می‌کنیم. سپس به معرفی و بررسی روش‌های رگرسیون استوار و رگرسیون تاوانیده به عنوان راهکارهای حل این مشکلات می‌پردازیم. همچنین با در نظر گرفتن مشاهده‌های دورافتاده و هم‌خطی چندگانه و یا مدل‌های تنک به‌طور هم‌زمان به بررسی روش‌های رگرسیون استوار تاوانیده می‌پردازیم.

در نهایت به‌منظور مقایسهٔ عملکرد براوردگرهای مختلف مطرح شده در این مقاله، ابتدا سه مطالعهٔ شبیه‌سازی را انجام داده و سپس به تحلیل یک مجموعه دادهٔ واقعی با استفاده از روش‌های رگرسیون استوار تاوانیده می‌پردازیم.

واژه‌های کلیدی: مشاهده‌های دورافتاده، رگرسیون استوار، هم‌خطی چندگانه، مدل تنک، رگرسیون تاوانیده.

۱ مقدمه

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 براورد رگرسیون خطی منسوب به گاوس و لژاندر است، که به روش کمترین توان‌های دوم معروف است. یکی از عواملی که براوردگر کمترین توان‌های دوم را تحت تأثیر قرار می‌دهد وجود مشاهده‌های دورافتاده در داده‌ها است، که از این تأثیر به عنوان عدم استواری براوردگر یاد می‌شود. لذا یافتن روش‌هایی که تحت تأثیر این نوع مشاهده‌های قرار نگیرند امری ضروری به حساب می‌آید، که این روش‌ها را رگرسیون استوار می‌نامند. از جمله روش‌های رگرسیون استوار می‌توان به براوردگر رگرسیونی کمترین قدر مطلق انحرافات [۴]، M-براوردگرها [۹]، کمترین میانهٔ توان‌های دوم [۱۴]، کمترین توان‌های دوم پیراسته [۱۵] و S-براوردگرها [۱۷] اشاره کرد.

علاوه بر وجود مشاهده‌های دورافتاده، وجود وابستگی خطی

[۱] دانش‌آموختهٔ کارشناسی ارشد آمار، دانشگاه تهران، ایران

[۲] عضو هیئت علمی گروه آمار، دانشگاه تهران، ایران

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

طرح کلّی این مقاله به‌این صورت است. در بخش دوم، پس از بیان مقدمات رگرسیون خطی و روش کلاسیک کمترین توان‌های دوم، به معرفی مشاهده‌های دورافتاده و نقاط نافذ و تأثیرات آنها بر این براوردگر کلاسیک رگرسیونی می‌پردازیم. سپس انواع براوردگرهای استوار را معرفی می‌کنیم. در بخش سوم، پس از معرفی هم‌خطی چندگانه، مشکل تعداد زیاد متغیرها در مقابل اندازهٔ کم نمونه و مدل‌های تنک و بررسی تأثیرات این مشکلات بر روش‌های رگرسیونی کلاسیک، به معرفی روش‌های رگرسیونی تاوانیده می‌پردازیم. در بخش چهارم، روش رگرسیون استوار کمترین توان‌های دوم پیراستهٔ تاوانیده بر اساس سه نوع متداول تابع تاوان را به عنوان راهکاری برای مقابلهٔ هم‌زمان با مشکلات هم‌خطی چندگانه و یا مدل‌های تنک و مشاهده‌های دورافتاده معرفی می‌کنیم. در بخش پنجم، روش رگرسیون استوار کمترین قدر مطلق انحرافات تاوانیده بر اساس سه نوع متداول تابع تاوان را به عنوان راهکاری دیگر برای مقابلهٔ هم‌زمان با مشکلات هم‌خطی چندگانه و یا مدل‌های تنک و مشاهده‌های دورافتاده بررسی می‌کنیم. در بخش ششم، به‌منظور مقایسهٔ عملکرد براوردگرهای مختلف مطرح شده در این مقاله، ابتدا سه مطالعهٔ شبیه‌سازی را انجام داده و سپس به تحلیل یک مجموعه دادهٔ واقعی با استفاده از روش‌های رگرسیون استوار تاوانیده می‌پردازیم.

۲ مشاهده‌های دورافتاده و روش‌های رگرسیون استوار

مدل رگرسیونی خطی به‌صورت

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

را در نظر بگیرید. که در آن $\mathbf{y}$ بردار $n \times ۱$ مشاهده‌های متغیر پاسخ،

$$\mathbf{X} = [\mathbf{X}_۱ \dots \mathbf{X}_p] = \begin{bmatrix} \mathbf{x}_۱ \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = (x_{ij})_{۱\le i\le n, ۱\le j\le p}$$

ماتریس $n \times p$ مشاهده‌های متغیرهای پیشگو، $\boldsymbol{\beta}$ بردار $p \times ۱$ ضرایب نامعلوم رگرسیونی و ε بردار $n \times ۱$ خطاها با میانگین صفر و ماتریس واریانس کواریانس $\sigma^۲\mathbf{I}$ است. هدف براورد بردار پارامترهای نامعلوم $\boldsymbol{\beta}$ از مجموعه مشاهده‌های نمونهٔ $(\mathbf{y}, \mathbf{X})$ است. یکی از متداول‌ترین روش‌های براورد بردار $\boldsymbol{\beta}$ روش کمترین توان‌های دوم است. برای مدل رگرسیونی خطی **(؟؟)** براوردگر رگرسیونی کمترین توان‌های دوم [۳] (LS) به‌صورت

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=۱}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^۲ . \qquad (۲)$$

به دست می‌آید. رابطهٔ **(؟؟)** به آسانی براوردگر

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \left(\mathbf{X}'\mathbf{X}\right)^{-۱}\mathbf{X}'\mathbf{y}. \qquad (۳)$$

را نتیجه می‌دهد. یکی از عواملی که براوردهای رگرسیونی کمترین توان‌های دوم را تحت تأثیر قرار می‌دهد وجود مشاهده‌های دورافتاده[۴] در داده‌ها است. دو نوع اصلی مشاهده‌های دورافتاده مشاهده‌های دورافتادهٔ عمودی و نقاط نافذ هستند که به‌صورت زیر تعریف می‌گردند. مشاهدهٔ $(y_k, \mathbf{x}_k)$ که در آن مقدار y_k دور از خط رگرسیونی‌ای که اکثر نقاط مجموعهٔ داده از آن تبعیت می‌کنند قرار دارد یک مشاهدهٔ دورافتادهٔ عمودی نامیده می‌شود. مشاهدهٔ $(y_k, \mathbf{x}_k)$ که در آن $\mathbf{x}_k$ دور از مجموعهٔ اکثریت $\mathbf{x}_i$ های مشاهده شده در

[۳] Least Squares

[۴] Outliers

نمونه قرار دارد یک نقطهٔ نافذ[۵] نامیده می‌شود. یکی از راه‌های تشخیص مشاهده‌های دورافتاده نمودارهای جزئی نقاط نافذ[۶] است که به‌صورت زیر تعریف می‌گردد.

فرض کنید $\mathbf{X}_{[k]}$ ماتریس $n \times (p-1)$ حاصل از حذف k امین ستون $\mathbf{X}$ ($\mathbf{X}_k$) و $\mathbf{u}_k$ و $\mathbf{v}_k$ به ترتیب مانده‌های حاصل از رگرسیون $\mathbf{y}$ روی $\mathbf{X}_{[k]}$ و رگرسیون $\mathbf{X}_k$ روی $\mathbf{X}_{[k]}$ باشند. در این صورت k امین ضریب رگرسیونی $\mathbf{y}$ روی $\mathbf{X}$ می‌تواند از رگرسیون ساده‌ی $\mathbf{u}_k$ روی $\mathbf{v}_k$ تعیین شود. نمودار جزئی نقاط نافذ برای $\hat{\boldsymbol{\beta}}_k$ به‌صورت نمودار پراکندگی $\mathbf{u}_k$ در برابر $\mathbf{v}_k$ به همراه خط رگرسیونی سادهٔ برازش داده شده به آنها تعریف می‌شود. به‌این ترتیب می‌توان مشاهده‌های دورافتادهٔ 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 براوردگرهایی با کارایی نسبتاً مطلوب نیز می‌شوند.

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

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

M-براوردگرها با

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(y_i - \mathbf{x}_i\boldsymbol{\beta}\right),$$

که در آن $\rho(\cdot)$ تابعی محدب، متقارن حول صفر و با مینیمم یکتا در صفر است، براوردگر کمترین میانهٔ توان‌های دوم [۹] (LMS) با

$$\hat{\boldsymbol{\beta}}_{\mathrm{LMS}} = \arg\min_{\boldsymbol{\beta}} Med_{1\le i\le n}\left(y_i - \mathbf{x}_i\boldsymbol{\beta}\right)^2,$$

براوردگر کمترین توان‌های دوم پیراسته [۱۰] (LTS) با

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

که در آن

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2_{1:n} \le \ldots \le (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2_{n:n}$$

مرتب شدهٔ توان دوم مانده‌ها و

$$h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right], \quad h = [0.75n] \qquad (۵)$$

مقادیر توصیه شده‌ای برای h هستند و S-براوردگرها با

$$\hat{\boldsymbol{\beta}}_{\mathrm{S}} = \arg\min_{\boldsymbol{\beta}} S\left(y_1 - \mathbf{x}_1\boldsymbol{\beta}, \ldots, y_n - \mathbf{x}_n\boldsymbol{\beta}\right),$$

که در آن $s = S\left(y_1 - \mathbf{x}_1\boldsymbol{\beta}, \ldots, y_n - \mathbf{x}_n\boldsymbol{\beta}\right)$ از معادلهٔ

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

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

به علت عدم استواری براوردگر رگرسیونی LAD در مقابل نقاط نافذ نقطهٔ فروریزش مجانبی این براوردگر به بالاتر از

[۵] Leverage Point

[۶] Partial Leverage Plots

[۷] Robust

[۸] Least Absolute Deviations

[۹] Least Median of Squares

[۱۰] Least Trimmed Squares

صفر درصد نمی‌رسد. در M-براوردگرها 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 براوردگر به دست آورد. نقطهٔ فروریزش مجانبی براوردگر LMS، LTS و S-براوردگرها (تحت شرایطی خاص) برابر با ۵۰ درصد است که بیشترین میزان نقطهٔ فروریزش برای یک براوردگر است [۱۶].

۳ هم‌خطی چندگانه، مدل‌های تنک و روش‌های رگرسیون تاوانیده

یکی دیگر از مشکلاتی که می‌تواند براورد 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 به‌صورت

$$VIF_j = \frac{۱}{۱ - R_j^۲},$$

محاسبه می‌شود. که در آن $R_j^۲$ ضریب تعیین[۱۳] چندگانهٔ رگرسیون $\mathbf{X}_j$ نسبت به دیگر متغیرهای پیشگوی مدل رگرسیونی است. عامل تورم واریانس بیش از ۱۰ نشان دهندهٔ مشکلات جدی هم‌خطی چندگانه است.

هم‌خطی چندگانه باعث می‌شود دترمینان ماتریس $\mathbf{X}'\mathbf{X}$ بسیار به صفر نزدیک شود و در نتیجه $\left(\mathbf{X}'\mathbf{X}\right)^{-۱}$ مقادیر بسیار بزرگی را اختیار کند. از آن‌جا که $Cov\left(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right) = \sigma^۲\left(\mathbf{X}'\mathbf{X}\right)^{-۱}$، این مسئله باعث تورم ماتریس واریانس کواریانس براوردگرهای کمترین توان‌های دوم ضرایب شده و استنباط‌های مربوط به ضرایب مدل رگرسیونی خطی را نامعتبر می‌سازد. یک راهکار برای کاهش این تورم انقباض و کشیدن براورد به سمت صفر است که باعث کاهش واریانس براوردگر 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 براورد ضرایب رگرسیونی ناصفر از اهمیت خاصی برخوردار است و مسئلهٔ انتخاب متغیر مطرح می‌گردد. در مدل‌های دارای متغیرهای پیشگوی بسیار زیاد و یا مدل‌های تنک انقباض همراه با انتخاب متغیر می‌تواند به براوردهای مناسبی منجر شود. منظم سازی[۱۵] یا رگرسیون تاوانیده[۱۶] راهکاری برای ایجاد انقباض در براوردگرهای رگرسیونی است که با اضافه کردن یک جملهٔ تاوان به تابع هدف مدل رگرسیونی صورت می‌گیرد. برای مدل رگرسیونی با تابع هدف $Q(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta})$ با اضافه کردن تابع تاوان $P(\boldsymbol{\beta})$، براوردگر رگرسیونی تاوانیده‌ی $\boldsymbol{\beta}$ به‌صورت

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

به دست می‌آید. تابع تاوان $P(\boldsymbol{\beta})$ یک اندازه برای میزان بزرگی $\boldsymbol{\beta}$ و پارامتر $\lambda > ۰$ برای تنظیم میزان تأثیر تابع تاوان در محدود سازی براورد حاصل است. بنا به انتخاب تابع تاوان 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.

[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 [۱۷] [۶ و ۷] یکی از متداول‌ترین تکنیک‌هایی است که در حالت هم‌خطی چندگانه مورد استفاده قرار می‌گیرد. در این روش تابع تاوان توان دوم نرم $L_۲$ی $\boldsymbol{\beta}$ یعنی $\boldsymbol{\beta}'\boldsymbol{\beta}$ است. براوردگر ستیغی به‌صورت

$$\hat{\boldsymbol{\beta}}_{\text{RIDGE}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=۱}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^۲ + \lambda \sum_{j=۱}^{p} \beta_j^۲ \right\}, \tag{۶}$$

به دست می‌آید. که در آن عدد حقیقی نامنفی λ میزان بزرگی تابع تاوان را تنظیم می‌کند. برای این براوردگر داریم

$$\hat{\boldsymbol{\beta}}_{\text{RIDGE}} = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-۱} \mathbf{X}'\mathbf{y}. \tag{۷}$$

به بیان دیگر براوردگر ستیغی معادل است با مسئلهٔ مینیمم سازی $\text{minimize}_{\boldsymbol{\beta}} \sum_{i=۱}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^۲$ تحت قید $\sum_{j=۱}^{p} \beta_j^۲ \leq M$، که در آن M یک کران ثابت است. در حقیقت یک رابطهٔ یک به یک بین پارامتر λ در **(؟؟)** و M وجود دارد.

براوردگر **(؟؟)** و **(؟؟)** تنها در یک جملهٔ $\lambda\mathbf{I}$ که به ماتریس $\mathbf{X}'\mathbf{X}$ اضافه شده است اختلاف دارند. اضافه شدن مقدار مثبت λ به مؤلفه‌های قطری ماتریس $\mathbf{X}'\mathbf{X}$ باعث کاهش تورم معکوس آن می‌شود.

[۸] پیشنهاد کرده‌اند که پارامتر ستیغی λ باید به اندازه‌ای کوچک انتخاب شود که

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

و بر این اساس مقدار بهینهٔ $\lambda = \frac{p\sigma^۲}{\boldsymbol{\beta}'\boldsymbol{\beta}}$ راپیشنهاد داده‌اند، که پارامترهای $\boldsymbol{\beta}$ و $\sigma^۲$ در آن رامی‌توان با $\hat{\boldsymbol{\beta}}_{\text{LS}}$ و $\hat{\sigma}^۲_{\text{LS}}$ جایگزین کرد.

۲.۳ روش کمترین قدر مطلق انقباض و عملگر انتخاب

یکی دیگر از روش‌های 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 $L_۱$ بردار $\boldsymbol{\beta}$ است. براوردگر LASSO به‌صورت

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=۱}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^۲ + n\lambda \sum_{j=۱}^{p} |\beta_j| \right\}, \tag{۸}$$

به دست می‌آید. با توجه به مشتق ناپذیر بودن تابع تاوان $\sum_{j=۱}^{p} |\beta_j|$ در نقطهٔ صفر، مقادیر بزرگ‌تری از λ منجر به صفر شدن ضرایب رگرسیونی متغیرهای پیشگوی کم تأثیر در مدل می‌گردد و در نتیجه عمل براورد پارامتر و انتخاب متغیر به‌صورت هم‌زمان انجام می‌گیرد.

به دلیل آنکه LASSO پارامترهای تنظیم یکسانی را برای همهٔ ضرایب رگرسیونی به کار می‌برد، براوردگرهای حاصل دارای اریبی قابل ملاحظه‌ای هستند. در راهکاری دیگر می‌توانیم با اختصاص دادن پارامترهای تنظیم متفاوت، تاوان کمتری برای ضرایب رگرسیونی متناظر با متغیرهای مهم در نظر گرفته و از اریبی براوردگرهای حاصل بکاهیم. براورد اصلاح شده‌ی LASSO که با نماد LASSO^* نمایش داده می‌شود به‌صورت

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}^*} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=۱}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^۲ + n \sum_{j=۱}^{p} \lambda_j |\beta_j| \right\}.$$

به دست می‌آید. یک الگوریتم سریع به‌منظور محاسبهٔ 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. نکتهٔ قابل توجه این است که رابطه‌ای مستقیم بین پارامتر تنظیم روش LASSO و تعداد گام‌های روش LARS وجود دارد.

[۱۸] Least Absolute Shrinkage and Selection Operator

[۱۹] Least Angle Regression

پارامتر تنظیم λی (؟؟) به گونه‌ای براورد می‌شود که یک معیار دقت براورد رگرسیونی را حد اقل کند. معیارهای مختلفی برای سنجش دقت براورد رگرسیونی وجود دارد که از جملهٔ 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 این‌که روش LASSO در بسیاری از تحلیل‌ها کارایی خوبی دارد، با این حال دارای محدودیت‌هایی است، که از جملهٔ آنها می‌توان به موارد زیر اشاره کرد:

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

م.۲) 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 انگارانهٔ[۲۴] این روش را بیان می‌کنیم.

براوردگر ساده انگارانهٔ تور مرتجع $\boldsymbol{\beta}$ به‌صورت

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

به دست می‌آید. نحوهٔ محاسبهٔ براوردگر تور مرتجع به این صورت است که با استفاده از داده‌های ساختگی می‌توان مسئلهٔ تور مرتجع ساده انگارانه را به یک مسئلهٔ LASSO تبدیل کرد [۲۰]. در حقیقت اندازهٔ نمونه در مسئلهٔ ساختگی $n+p$ است، 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 تورمرتجع در غلبه بر م.۲ از روش LASSO محدب بودن اکید تابع تاوان روش تور مرتجع نسبت به $\boldsymbol{\beta}$، به‌ازای $\lambda_2 > 0$ است [۲۰].

عملکرد براوردگر تور مرتجع ساده انگارانه در دو مرحله صورت می‌گیرد، ابتدا برای هر λ_2ی ثابت یک براورد رگرسیونی ستیغی به دست آورده و سپس این براورد 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].

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

مسئلهٔ مهم دیگر در روش تور مرتجع انتخاب مقادیر مناسب برای پارامترهای تنظیم (λ_1, λ_2) است. همان‌طور که در مورد براوردگر ستیغی بیان کردیم رابطه‌ای مستقیم بین پارامتر تنظیم λی براورد ستیغی با کران M در قید $\sum_{j=1}^{p} \beta_j^2 \le M$ 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.

[۲۰] Akaike Information Criterion

[۲۱] Bayesian Information Criterion

[۲۲] Cross Validation

[۲۳] Elastic Net

[۲۴] Naive

۴ براوردگر کمترین توان‌های دوم پیراستهٔ تاوانیده

روش‌های رگرسیونی تاوانیدهٔ 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 شیوهٔ براورد 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 پیراستهٔ ستیغی

روش کمترین توان‌های دوم پیراستهٔ 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 جملهٔ تاوان توان دوم نرم L_2ی $\boldsymbol{\beta}$ با پارامتر تنظیم λ به تابع هدف کمترین توان‌های دوم پیراسته به‌صورت

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

به دست می‌آید. که در آن

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

و h به ترتیب در رابطه‌های (؟؟) و (؟؟) داده شده‌اند.
به بیان دیگر می‌توان نوشت

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

و در نتیجه

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

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

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

و

$$Q(\mathbf{z}) = \left(\mathbf{y} - \mathbf{X} \left(\mathbf{X}' \mathbf{Z} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}' \mathbf{Z} \mathbf{y} \right)' \mathbf{Z} \left(\mathbf{y} - \mathbf{X} \left(\mathbf{X}' \mathbf{Z} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}' \mathbf{Z} \mathbf{y} \right) + \lambda \mathbf{y}' \mathbf{Z}' \mathbf{X} \left(\mathbf{X}' \mathbf{Z} \mathbf{X} + \lambda \mathbf{I} \right)^{-2} \mathbf{X}' \mathbf{Z} \mathbf{y}.$$

به‌منظور محاسبهٔ براورد $\mathrm{LTS-RIDGE}$ الگوریتم $\mathrm{FAST-LTS-RIDGE}$ را به کار می‌بریم که به‌صورت زیر است.
برای یک پارامتر تنظیم ثابت λ تابع هدف مجموع توان‌های دوم تاوانیده با توان دوم نرم L_2ی $\boldsymbol{\beta}$ روی زیرمجموعهٔ $\mathbf{H} \subseteq \{1, \ldots, n\}$ با $|\mathbf{H}| = h$ را به‌صورت

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

در نظر بگیرید. برای انتخاب زیرمجموعهٔ اولیه $(\mathbf{H}_\cdot)$ به این صورت عمل می‌کنیم که ابتدا زیرمجموعه‌ای دلخواه از اندازه‌ی h انتخاب کرده و برای زیرمجموعهٔ مذکور پارامتر λ را به‌کمک معیار GCV براورد می‌کنیم. قرار می‌دهیم

$$\mathbf{H}_1 = \left\{ (y_i, \mathbf{x}_i) ; e^2_{\cdot,i} \in \left\{ (\mathbf{e}^2_{\cdot})_{j:n} : j = 1, \ldots, h \right\} \right\},$$

در مراحل بعدی الگوریتم برای براورد پارامتر تنظیم λ از روش براورد پارامتر ارائه شده توسط [۸] که در بخش قبل به آن اشاره کردیم استفاده می‌کنیم. در مرحله‌ی k ام بر اساس زیرمجموعهٔ جاری $\mathbf{H}_k$ که $|\mathbf{H}_k| = h$، ساختن زیرمجموعهٔ بعدی $\mathbf{H}_{k+1}$ از مشاهده‌های متناظر با h تا از کوچک‌ترین توان دوم مانده‌های حاصل از به کار گرفتن براوردگر ستیغی روی زیرمجموعهٔ $\mathbf{H}_k$ صورت می‌پذیرد. همچنین می‌توان با تکرار زیرمجموعهٔ ابتدایی به یک مینیمم موضعی تا حد امکان نزدیک به مینیمم مطلق دست یافت.

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

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

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

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

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

محاسبه شود.

۲.۴ روش رگرسیونی کمترین توان‌های دوم پیراستهٔ تنک

با وجود استواری بالای روش LTS واضح است که این روش براوردهای مدل تنک را تولید نمی‌کند. یک نسخهٔ تنک از LTS با افزودن یک تابع تاوان نرم L_1 بردار $\boldsymbol{\beta}$ با پارامتر تنظیم λ به تابع هدف LTS حاصل می‌شود. براوردگر کمترین توان‌های دوم پیراستهٔ تنک (Sparse − LTS) به‌صورت

$$\boldsymbol{\beta}_{\mathrm{Sparse-LTS}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{h} \left((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2\right)_{i:n} + h\lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

به دست می‌آید. که در آن

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2_{1:n} \le \ldots \le (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2_{n:n}$$

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

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

نقطهٔ فروریزش براوردگر رگرسیونی LTS − LASSO $\frac{n-h+1}{n}$ 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 براوردگر 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 برمحدودیت‌های ناشی از روش LASSO که در زیربخش قبل بیان کردیم، [۱۳] روش رگرسیونی کمترین توان‌های دوم پیراستهٔ تور مرتجع را برای تعدیل مشکلات مشاهده‌های دورافتاده و هم‌خطی چندگانه مطرح کردند. براوردگر رگرسیونی کمترین توان‌های دوم پیراستهٔ تور مرتجع ساده انگارانه (LTS − N − EN) به‌صورت

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

به دست می‌آید. که در آن

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2_{1:n} \le \ldots \le (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2_{n:n}$$

و h به ترتیب در رابطه‌های **(؟؟)** و **(؟؟)** داده شده‌اند.
با توجه به این‌که روی هر زیرمجموعهٔ h تایی براوردگر LTS − N − EN تبدیل به یک براوردگر N − EN می‌شود، بنا بر این با استفاده از داده‌های ساختگی روی هر زیرمجموعهٔ h تایی براوردگر N − EN تبدیل به یک براوردگر 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 براوردگر LTS − N − EN دست یافت.

۵ رگرسیون استوار در حضور هم‌خطی چندگانه به وسیلهٔ برآورگر کمترین قدر مطلق انحرافات تاوانیده

اضافه کردن یک جملهٔ تاوان به تابع هدف براوردگر LAD می‌تواند منجر به براوردگرهای رگرسیونی استواری شود که در حضور هم‌خطی چندگانه و یا در مواجهه با مدل‌های تنک نیز دارای کارایی و عملکرد پیشگویی قابل اطمینانی هستند. در این بخش به بررسی براوردگرهای LADی تاوانیده توسط تابع تاوان نرم L_1 و توان دوم نرم L_2ی بردار ضرایب رگرسیونی و ترکیبی خطی از این دو نرم می‌پردازیم.

۱.۵ روش کمترین قدر مطلق انحرافات ستیغی

ایدهٔ اصلی روش کمترین قدر مطلق انحرافات ستیغی (LAD − RIDGE) افزودن یک تابع تاوان توان دوم نرم L_2ی بردار ضرایب (RIDGE) به تابع هدف کمترین قدر مطلق انحرافات (LAD) است. این روش در مقایسه با روش کمترین قدر مطلق انحرافات می‌تواند عمل انقباض ضرایب رگرسیونی و براورد پارامتر را به‌صورت هم‌زمان انجام دهد و در مقایسه با براوردگر ستیغی در مقابل مشاهده‌های دورافتادهٔ عمودی استوار است. براوردگر LAD − RIDGE به‌صورت

$$\hat{\boldsymbol{\beta}}_{\text{LAD-RIDGE}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} |y_i - \mathbf{x}_i\boldsymbol{\beta}| + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \quad (۹)$$

به دست می‌آید. به‌منظور محاسبهٔ براوردگر LAD − RIDGE از تقریب خطی موضعی بر اساس بسط تیلور به‌صورت

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

استفاده می‌کنیم. که در آن $\boldsymbol{\beta}_j^{(\cdot)}$ یک مقدار معلوم است. در نتیجه براوردگر LAD − RIDGE را می‌توان یک براوردگر LAD − LASSO* تقریبی در نظر گرفت، که در زیر بخش بعد به معرفی آن می‌پردازیم.

۲.۵ روش کمترین قدر مطلق انحرافات تنک

با افزودن یک تابع تاوان نرم L_1 بردار ضرایب رگرسیونی به تابع هدف LAD می‌توان از این روش در مواجهه با مدل‌های تنک بهره گرفت. این روش در مقایسه با LAD می‌تواند عمل انتخاب متغیر و براورد پارامتر را به‌صورت هم‌زمان انجام دهد و در مقایسه با LASSO در مقابل مشاهده‌های دورافتادهٔ عمودی استوار است.

همان‌طور که در بخش سوم اشاره کردیم، نسخهٔ اصلاح شده‌ی LASSO، که آن را با نماد LASSO* نشان دادیم، دارای کارایی بالاتری نسبت به LASSOی معمولی است. علیرغم پیچیدگی بیشتر LASSO* نسبت به LASSO، ترجیح می‌دهیم از ترکیب دو روش LAD و LASSO* استفاده کنیم. براوردگر LAD − LASSO* به‌صورت

$$\hat{\boldsymbol{\beta}}_{\text{LAD-LASSO}^*} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} |y_i - \mathbf{x}_i\boldsymbol{\beta}| + n \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}. \quad (۱۰)$$

به دست می‌آید. از نظر محاسباتی مسئلهٔ یافتن براوردگر LAD − LASSO* با استفاده از داده‌های ساختگی به یک مسئلهٔ یافتن یک براوردگر LAD کاهش می‌یابد. به‌منظور محاسبهٔ براوردگر LAD می‌توان از الگوریتم BR [۳] استفاده کرد.

به دلیل وجود p پارامتر تنظیم در مدل (؟؟) و براورد هم‌زمان این پارامترها، استفاده از روش‌های CV، GCV و BIC برای براورد این پارامترها مناسب نیست. بر اساس ایده‌ای از [۱۹] براوردگر LAD − LASSO* را می‌توان به عنوان یک براوردگر بیزی در نظر گرفت، به‌طوری که هر ضریب رگرسیونی β_j دارای توزیع پیشین نمایی دوگانه با پارامتر مکانی صفر و پارامتر مقیاسی $n\lambda_j$ است و بر این اساس مقدار بهینه‌ی $\hat{\lambda}_j = \frac{1}{n|\beta_j|}$ را ارئه شده است، که در آن β_j را می‌توان با استفاده از براوردگر LADی معمولی براورد کرد.

۳.۵ روش کمترین قدر مطلق انحرافات تور مرتجع

ایدهٔ 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 براورد پارامتر را به‌صورت هم‌زمان انجام دهد و در مقایسه با براوردگر تور مرتجع در مقابل مشاهده‌های دورافتادهٔ عمودی استوار است. همچنین این روش معایب روش LAD − LASSO در انتخاب متغیر را ندارد. براوردگر کمترین قدر مطلق انحرافات تور مرتجع ساده انگارانه به‌صورت

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

به دست می‌آید. به‌منظور سهولت محاسبهٔ براوردگر LAD − N − EN، از تقریب

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

استفاده می‌کنیم. در نتیجه براوردگر LAD − N − EN به یک براوردگر LAD − LASSO* تقریبی تبدیل می‌شود.

۶ مطالعهٔ عددی، شبیه‌سازی و بررسی داده‌های واقعی

به‌منظور مقایسهٔ عملکرد براوردگرهای مختلف مطرح شده در بخش‌های چهارم و پنجم ابتدا سه مطالعهٔ شبیه‌سازی مجزا را انجام می‌دهیم. سپس به تحلیل یک مجموعه دادهٔ واقعی چربی بدن [۱۰] با استفاده از روش‌های رگرسیون استوار تاوانیده‌ی LTS − LASSO و LTS − RIDGE می‌پردازیم.

۱.۶ شبیه‌سازی

در این بخش به مطالعهٔ سه شبیه‌سازی مجزا برای مقایسهٔ عملکرد براوردگرهای مختلف مطرح شده در بخش‌های چهارم و پنجم می‌پردازیم. این شبیه‌سازی‌ها در نرم‌افزار R انجام شده است. هدف از شبیه‌سازی اول مقایسهٔ عملکرد روش‌های LASSO و EN برای تنک سازی برآوردگرهای به‌دست‌آمده از یک طرف و همچنین مقایسهٔ روش‌های LAD و LTS برای استوار سازی براوردهای حاصل در برابر مشاهده‌های دورافتاده از طرف دیگر است. به‌این منظور چهار ترکیب مختلف از این روش‌ها را به‌صورت LAD − LASSO*، LTS − LASSO، LAD − EN و LTS − EN در نظر می‌گیریم. علاوه بر محاسبه و مقایسهٔ عملکرد این چهار براوردگر، دو روش LTS و LAD را نیز به‌منظور بررسی میزان تأثیرگذاری توابع تاوان در بهبود براوردگرهای حاصل و دو روش LASSO و EN را نیز به‌منظور بررسی روش‌های استوار سازی در دستور کار قرار می‌دهیم. همچنین عملکرد براورد کمترین توان‌های دوم (LS) را نیز در کنار براوردگرهای بالا بررسی می‌کنیم. ملاک‌های بررسی دقت عملکرد براوردگرها در این مطالعهٔ شبیه‌سازی، ملاک ریشهٔ 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 به‌صورت زیر تعریف می‌گردند.

ریشهٔ دوم میانگین توان‌های دوم خطای پیشگویی (**RMSPE**): به‌منظور محاسبهٔ این ملاک، ابتدا n مشاهدهٔ اضافی (y_i^*, x_i^*)، $i = 1, \ldots, n$، از مدل مورد نظر، بدون مشاهده‌های دورافتاده، به عنوان دادهٔ آزمایشی تولید می‌کنیم.

[۲۵] Root Mean Squared Prediction Error

[۲۶] False Positive Rate

[۲۷] False Negative Rate

حال ملاک RMSPE را به‌صورت

$$\mathrm{RMSPE}\left(\hat{\boldsymbol{\beta}}\right)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^*-\mathbf{x}_i^*\hat{\boldsymbol{\beta}}\right)^2},$$

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

$$\mathrm{FPR}\left(\hat{\boldsymbol{\beta}}\right)=\frac{\left|\left\{j\in\{1,\ldots,p\}:\hat{\beta}_j\neq 0,\beta_j=0\right\}\right|}{\left|\{j\in\{1,\ldots,p\}:\beta_j=0\}\right|}.$$

به دست می‌آید. نرخ منفی نادرست (**FNR**): این ملاک در واقع نسبت ضرایبی است که به اشتباه براورد آنها صفر شده است و به‌صورت

$$\mathrm{FNR}\left(\hat{\boldsymbol{\beta}}\right)=\frac{\left|\left\{j\in\{1,\ldots,p\}:\hat{\beta}_j=0,\beta_j\neq 0\right\}\right|}{\left|\{j\in\{1,\ldots,p\}:\beta_j\neq 0\}\right|}.$$

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

هدف از شبیه‌سازی دوم مقایسهٔ عملکرد روش‌های براورد ستیغی استوار مختلف است. به‌این منظور سه روش $LAD-RIDGE$، $LTS-RIDGE$ ساده انگارانه و $LTS-RIDGE$ دقیق را مورد مطالعه و مقایسه قرار می‌دهیم. علاوه بر محاسبه و مقایسهٔ عملکرد این سه براوردگر، دو روش LTS و LAD را نیز به‌منظور بررسی میزان تأثیرگذاری اریب سازی ستیغی در بهبود براوردهای حاصل و روش RIDGE را نیز به‌منظور بررسی روش‌های استوار سازی در دستور کار قرار می‌دهیم. همچنین عملکرد براورد کمترین توان‌های دوم (LS) را نیز در کنار براوردگرهای بالا بررسی می‌کنیم. در این شبیه‌سازی تنها ملاک بررسی دقت عملکرد براوردگرها، ریشهٔ دوم میانگین توان‌های دوم خطای پیشگویی است. هدف از شبیه‌سازی سوم مقایسهٔ همهٔ روش‌های مطرح شده در دو شبیه‌سازی قبل در مواجهه با یک مدل تنک و حضور هم‌خطی چندگانه است. ملاک‌های بررسی دقت عملکرد براوردگرها در این مطالعهٔ شبیه‌سازی نیز مشابه شبیه‌سازی است.

۱.۱.۶ شبیه‌سازی اول

در هر بار تکرار این شبیه‌سازی تعداد $n=100$ مشاهدهٔ متغیرهای پیشگو از یک توزیع نرمال p متغیره‌ی $N_p(0,\mathbf{I}_{p\times p})$ تولید شده است و بردار ضرایب رگرسیونی به‌صورت $\boldsymbol{\beta}=(10,0,0,15,0)$ داده شده است. در نتیجه در این شبیه‌سازی با یک مدل تنک مواجه هستیم. متغیرهای پاسخ طبق مدل $y_i=\mathbf{x}_i\boldsymbol{\beta}+\varepsilon_i$ تولید شده است، که در آن خطاهای ε_i از توزیع نرمال $N(0,0/5)$ تولید شده‌اند.

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

(۱). عدم ایجاد مشاهده‌های دورافتاده.

(۲). ایجاد مشاهده‌های دورافتادهٔ عمودی: ۱۰٪ از خطاهای مدل به جای توزیع $N(0,0/5)$ از توزیع $N(40,0/5)$ تولید شده‌اند.

(۳). ایجاد مشاهده‌های دورافتادهٔ 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 براوردگرهای مختلف محاسبه شده در این شبیه‌سازی و به ترتیب به‌ازای سه طرح مختلف «بدون دورافتاده» (طرح (۱))، «دورافتادهٔ عمودی» (طرح(۲)) و «دورافتادهٔ نافذ» (طرح(۳)) نشان می‌دهد.

همان‌طور که از جدول ۱ قابل مشاهده است، در صورت عدم وجود مشاهده‌های دورافتاده، با توجه به این‌که با یک مدل تنک مواجه هستیم، براوردگرهای EN و LASSO عملکرد پیشگویی خوبی را از دیدگاه RMSPE و FPR از خود نشان داده‌اند. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی براوردگرهای LS، LAD و LTS با توجه به کاهش میزان RMSPE و FPR براوردگرهای تاوانیده بر اساس این سه براوردگر مشهود 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 RMSPE ها 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
دورافتادهٔ عمودی	۴/۷۳۶	۰/۵۵۵	۰/۵۵۲	۴/۳۹۰	۰/۷۸۱	۰/۵۴۶	۰/۶۳۴
دورافتادهٔ نافذ	۰/۶۸۹۹	۰/۵۶۷	۰/۸۴۹	۵/۶۴۱	۰/۶۰۸	۰/۵۶۲	۰/۷۷۵

با توجه به جدول ۲ در صورت وجود مشاهده‌های دورافتادهٔ عمودی در مجموعه دادهٔ مورد بررسی، همان‌طور که انتظار داشتیم براوردگرهای $LAD - LASSO^*$ و $LTS - LASSO$ دارای عملکرد پیشگویی خوبی از دیدگاه RMSPE هستند. مقدار RMSPEی بالای دو روش LASSO و EN از عدم استواری این دو روش در مقابل مشاهده‌های دورافتادهٔ 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 براوردگرهای LS، LAD و LTS با توجه به کاهش میزان RMSPE و FPR براوردگرهای تاوانیده بر اساس این سه براوردگر مشهود است.

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

پیشگویی خوبی را از خود نشان نداده‌اند. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی براوردگرهای LS، LAD و LTS با توجه به کاهش میزان FPR براوردگرهای تاوانیده بر اساس این سه براوردگر 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 براورد 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 به‌صورت $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$ با $\sigma_{ij} = 0/99^{|i-j|}$ و بردار ضرایب رگرسیونی به‌صورت $\boldsymbol{\beta} = (10, 5, 8, 15, 4)$ 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 براوردگرهای 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 براوردگر RIDGE دارای کمترین مقدار RMSPE است و در رتبهٔ بعدی براوردگر LS قرار دارد. به‌طور کلّی با توجه به این‌که براورد پارامتر تنظیم براوردگر RIDGE به نحوی صورت می‌پذیرد که $\text{MSE}\left(\hat{\boldsymbol{\beta}}_{\text{RIDGE}}\right) \leq \text{MSE}\left(\hat{\boldsymbol{\beta}}_{\text{LS}}\right)$، 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 براوردگرهای LS، LAD و LTS با توجه به کاهش میزان RMSPEی براوردگرهای تاوانیده بر اساس این سه براوردگر مشهود است. در صورت وجود مشاهده‌های دورافتادهٔ عمودی همان‌طور قابل مشاهده است، دو براوردگر $\text{LAD} - \text{RIDGE}$ و $\text{LTS} - \text{RIDGE2}$ دارای بهترین عملکرد پیشگویی از دیدگاه RMSPE هستند. RMSPEی بالای دو براوردگر LS و RIDGE از عدم استواری این دو براوردگر در مقابل مشاهده‌های دورافتادهٔ عمودی ناشی شده است. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی براوردگرهای LS، LAD و LTS با توجه به کاهش میزان RMSPEی براوردگرهای تاوانیده بر اساس این سه براوردگر مشهود است. در صورت وجود هم‌زمان مشاهده‌های دورافتادهٔ عمودی و نقاط نافذ، همان‌طور که قابل مشاهده است، براوردگر $\text{LTS} - \text{RIDGE2}$ دارای بهترین عملکرد پیشگویی از دیدگاه RMSPE است. RMSPEی بالای براوردگرهای LS و RIDGE ناشی از عدم استواری این براوردگرها در مقابل مشاهده‌های دورافتادهٔ عمودی و نقاط نافذ است. همچنین اثر اریب سازی بر بهبود عملکرد پیشگویی براوردگرهای LS، LAD و LTS، با توجه به کاهش میزان RMSPEی براوردگرهای تاوانیده بر اساس این سه براوردگر مشهود است.

۳.۱.۶ شبیه‌سازی سوم

در هر بار تکرار این شبیه‌سازی نیز تعداد $n = 100$ مشاهدهٔ متغیرهای پیشگو از یک توزیع نرمال مشابه شبیه‌سازی دوم تولید شده است و بردار ضرایب رگرسیونی نیز مشابه شبیه‌سازی اول به‌صورت $\boldsymbol{\beta} = (10, 0, 0, 15, 0)$ 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 $y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$ 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 براوردگرهای مختلف محاسبه شده در این شبیه‌سازی و به ترتیب به‌ازای دو طرح مختلف «دورافتادهٔ عمودی» (طرح(۲)) و «دورافتادهٔ نافذ» (طرح(۳)) نشان می‌دهد.

همان‌طور که از جدول ۵ قابل مشاهده است، در صورت وجود مشاهده‌های دورافتادهٔ عمودی دو براوردگر $LAD-LASSO*$ و $LTS-LASSO$ دارای عملکرد پیشگویی خوبی از دیدگاه RMSPE هستند. RMSPEی بالای دو روش RIDGE و EN از عدم استواری این دو روش در مقابل مشاهده‌های دورافتادهٔ عمودی و نیز عدم قابلیت براوردگر RIDGE در تنک سازی است. RMSPEی 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 چندگانهٔ موجود در مدل مورد بررسی است و در نتیجه RMSPEی 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 درجدول ۶ قابل مشاهده است دو روش $LTS-LASSO$ و LTS دارای عملکرد پیشگویی خوبی از دیدگاه RMSPE هستند که شاید RMSPEی 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 براوردگر $LTS-LASSO$ به علت دارا بودن این قابلیت عملکر خوبی را از دیدگاه FPR و همچنین FNR از خود نشان داده است.

۲.۶ بررسی دادهٔ واقعی

این مثال مربوط به داده‌های واقعی چربی بدن است که حاصل از مطالعه‌ای به‌منظور براورد چگالی بدن ۲۵۲ مرد بر اساس اندازه‌گیری اندازه‌های مختلف اندام آنها صورت گرفته است. این مجموعه داده در بستهٔ 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، اندازهٔ دور سینه، اندازهٔ دور شکم، اندازهٔ دور مفصل ران، اندازهٔ دور ران، اندازهٔ دور زانو، اندازهٔ قوزک پا، اندازهٔ دو سر بازو، اندازهٔ دور بازو و اندازهٔ دور مچ هستند. مدل رگرسیونی خطی به‌صورت

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

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 به‌صورت زیر است.

با توجه به مقادیر VIF بالا به خصوص برای متغیرهای Weight، Chest، Abdomen و Hip می‌توان مشاهده کرد که هم‌خطی چندگانهٔ 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.

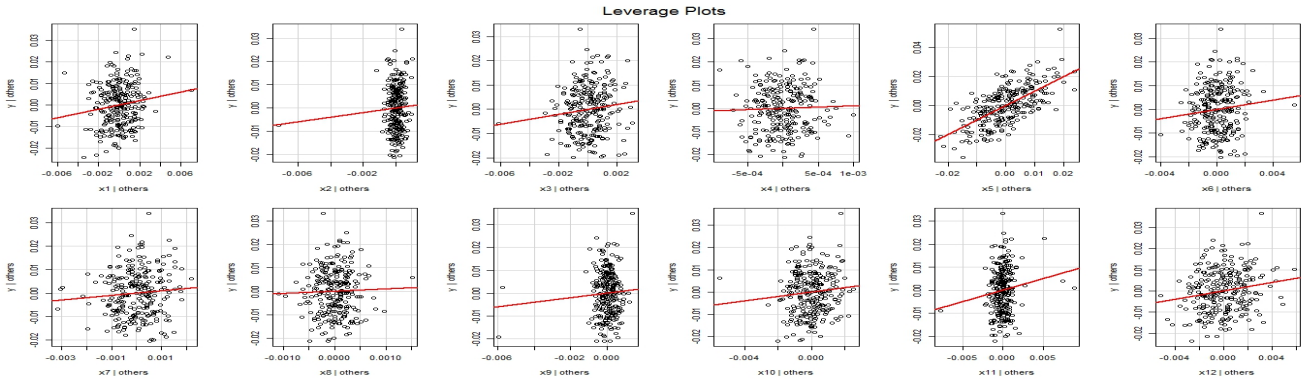
براورد ضرایب رگرسیونی به روش‌های $LTS-LASSO$ و $LTS-RIDGE$ دقیق به ترتیب به شرح زیر است

جدول ۵. میانگین RMSPE، FPR و FNRهای شبیه‌سازی 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 FNRهای شبیه‌سازی 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 نتیجه‌گیری

در این مقاله پس از بررسی نقایص براوردگر رگرسیونی کلاسیک کمترین توان‌های دوم در برابر مشاهده‌های دورافتاده، هم‌خطی چندگانه و مواجهه با مدل‌های تنک، به معرفی روش‌های رگرسیون 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 به‌منظور بهبود هر چه بیشتر عملکرد این براوردگرها و کاهش میزان اریبی و واریانس و افزایش کارایی آنها می‌توان توابع تاوان دیگری را در ساختار تابع هدف این براوردگرها به کار گرفت. توابع تاوان غیر مقعر و از جملهٔ مهم‌ترین و بهترین آنها تابع تاوان SCAD گزینهٔ مناسبی به‌این منظور است که در کنار یک روش رگرسیونی استوار با نقطهٔ فروریزش بالا براوردگرهای رگرسیونی استوار تاوانیدهٔ 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.

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.