مجله علوم آماری، پاییز و زمستان ۱۳۹۹
جلد ۱۴، شماره ۲، ص ۴۰۹ – ۴۲۸
DOI: 10.29252/jss.14.2.409

مدل‌سازی سن تقویمی به روش رگرسیون ستیغی کمترین توان‌های دوم پیراسته

مهدی روزبه، منیره معنوی

گروه آمار، دانشکده علوم ریاضی، آمار و علوم کامپوتر، دانشگاه سمنان

تاریخ دریافت: ۱۳۹۸/۰۲/۲۶ تاریخ آخرین بازنگری: ۱۳۹۸/۱۰/۳۰

چکیده: متداول‌ترین روش برای برآورد پارامترهای مدل رگرسیون خطی، روش کمترین توان‌های دوم معمولی است که علی‌رغم سادگی محاسبه و دستیابی به بهترین برآورد خطی ناریب از پارامترها، گاهی منجر به جواب‌های گمراه‌کننده می‌شود. به عنوان مثال می‌توان به مشکلات ناشی از وجود همخطی و داده‌های دورافتاده در مجموعه داده‌ها اشاره کرد. روش کمترین توان‌های دوم پیراسته که یکی از معروف‌ترین روش‌های رگرسیون استوار است، تاثیر داده‌های دورافتاده را تا حد امکان کم می‌کند. هدف اصلی این مقاله ارائه یک برآورد ستیغی استوار در مدل‌سازی مربوط به داده‌های سن دندانی است. در بین روش‌هایی که برای تعیین سن استفاده می‌شود، رایج‌ترین روش در سراسر دنیا، روش نوین تعمیم‌یافته دمیرجیان است که بر اساس سخت‌شدگی دندان دائمی در رادیوگرافی پانورامیک بنا شده است. نشان داده شده است که استفاده از برآوردگر ستیغی استوار منجر به کاهش میانگین توان دوم خطای برآورد در مقایسه با برآوردگر کمترین توان‌های دوم معمولی می‌شود. البته برآوردگرهای پیشنهادی در داده‌های شبیه‌سازی‌شده نیز مورد ارزیابی قرار گرفتند.

واژه‌های کلیدی: برآوردگر ستیغی استوار، سن دندانی دمیرجیان، کمترین توان‌های دوم پیراسته.

آدرس الکترونیکی نویسنده مسئول مقاله: مهدی روزبه، mahdi.roozbeh@semnan.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J07.

۱ مقدمه

مدل رگرسیونی خطی به‌صورت

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = ۱, \ldots, n,$$

است، که در آن y_i متغیر پاسخ، $\mathbf{x}_i^\top = (x_{i۱}, \ldots, x_{ip})$، بردار متغیر توضیحی، $\boldsymbol{\beta}$ بردار مجهول پارامترها و ε_i جمله خطا با شرایط

$$E(\varepsilon_i) = ۰, \quad Var(\varepsilon_i) = \sigma^۲, \quad Cov(\varepsilon_i, \varepsilon_j) = ۰, \ \ i \neq j,$$

است. برآوردگر بردار ضرایب رگرسیونی به روش کمترین توان‌های دوم به صورت $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-۱} \mathbf{X}^\top \mathbf{y}$ است، که در آن $\mathbf{X} = (\mathbf{x}_۱, \ldots, \mathbf{x}_n)^\top$ و $\mathbf{y} = (y_۱, \ldots, y_n)^\top$. این برآورد، بهترین برآوردگر نااریب خطی با کمترین واریانس[1] است. مانده‌ها بر اساس برآوردگر کمترین توان‌های دوم به‌صورت $e_i = y_i - \hat{y}_i$ تعریف می‌شود که در آن $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ مقدار پیش‌بینی شده برای مشاهده iام است. این روش بسیار ساده و کاربردی است اما در مواردی نظیر وجود نقطه پرت در مجموعه داده و همخطی بین متغیرهای توضیحی بسیار ضعیف عمل می‌کند.

۲ روش‌های استوار

اولین برآورد استوار توسط اجورث (۱۸۸۷) معرفی شد. وی استدلال کرد که نقاط دورافتاده تاثیر زیادی روی روش کمترین توان‌های دوم دارند. زیرا این روش در برآورد پارامترها به هر مشاهده وزن یکسان می‌دهد و بنابراین در نظرگرفتن توان دوم مانده‌ها باعث افزایش تاثیر نقاط دورافتاده روی برآوردگرها می‌شود. به عبارت دیگر، برآوردگر کمترین توان‌های دوم معمولی از کمینه‌سازی عبارت $\sum_{i=۱}^n e_i^۲$ حاصل می‌شود. اجورث (۱۸۸۷) پیشنهاد کرد، به جای استفاده از مسئله کمینه‌سازی کمترین توان‌های دوم عبارت $\sum_{i=۱}^n |e_i|$ کمینه شود. یعنی به جای استفاده از توان دوم مانده‌ها، از قدرمطلق آن‌ها استفاده شود. مارونا و همکاران (۲۰۰۶) معیار استواری یک برآوردگر را با استفاده از نقطه شکست[2] تعریف کردند.

[1] Best linear Unbiased Estimator (BLUE)
[2] Breakdown point

قدیمی‌ترین تعریف نقطه شکست را هاگ (۱۹۷۴) ارائه داد. البته همپل (۱۹۷۱) رابطه کاربردی‌تری تعریف کرد. دونهو و هوبر (۱۹۸۳) نیز تعریف ساده‌ای ارائه کردند که کاربردی‌ترین تعریف برای نقطه شکست است که در ادامه بیان می‌شود. البته قبل از آن نیاز به ارائه مفهوم نرم مرتبه دوم است.

تعریف ۱. *تابع حقیقی* $||\cdot||_2$ *نرم مرتبه دوم نامیده می‌شود، اگر به ازای هر بردار دلخواه* $\mathbf{a} = (a_1, \ldots, a_p)$ *رابطه* $||\mathbf{a}||_2^2 = (\sum_{i=1}^{p} |a_i|^2)^{\frac{1}{2}}$ *برقرار باشد.*

تعریف ۲. *(مارونا و همکاران ، ۲۰۰۶) نقطه شکست برآوردگر* $T = T(\mathbf{W})$ *برای نمونه* $\mathbf{W}$ *به حجم* n *به‌صورت* $\mathrm{BP}(T; \mathbf{W}) = \min_m\{m : \sup_{\mathbf{W}^*} ||T(\mathbf{W}^*)||_2^2 = \infty\}$ *تعریف می‌شود، که در آن* $\mathbf{W}^*$ *نمونه آلوده شده*[1] *با جایگزینی* $m \leq n$ *نقطه از نمونه* $\mathbf{W}$ *با مقادیر دلخواه است.*

بنابراین درصد شکست برآوردگر T برای یک نمونه متناهی مانند $\mathbf{W}$ برابر با $\frac{m}{n}$ خواهد بود. برای رگرسیون کمترین توان‌های دوم معمولی، وجود یک داده غیر عادی برای تحت تاثیر قراردادن برآورد ضرایب کافی است. بنابراین نقطه شکست آن به صورت $\mathrm{BP}(T(\mathbf{W}), \mathbf{W}) = \frac{1}{n}$ است. هر چه n بزرگتر باشد، $\frac{1}{n}$ به سمت صفر میل می‌کند، به این معنی که درصد نقطه شکست برآوردگر کمترین توان‌های دوم معمولی صفر درصد است. یعنی تنها حضور یک مشاهده دورافتاده می‌تواند تاثیر جدی و مخربی روی برآورد کمترین توان‌های دوم معمولی داشته باشد. روش‌های استوار مختلف دیگری برای غلبه بر آثار بد نقاط دورافتاده توسط محققان پیشنهاد شده است که از جمله آن‌ها می‌توان به روش‌های والد (۱۹۴۰)، نیر و سریواستاوا (۱۹۴۲)، بارتلت (۱۸۸۷)، براون وموود (۱۹۵۱)، توکی (۱۹۷۰)، جرکوا (۱۹۷۱)، جیگل (۱۹۷۲)، بیکل (۱۹۷۳)، اندروز (۱۹۷۴)، کونکر و باست (۱۹۷۸) و ولمن و باست (۱۹۸۱) اشاره نمود. متاسفانه درصد شکست این روش‌ها حداکثر ۳۰ درصد است و همچنین برخی از آن‌ها برای $p > 2$ (p تعداد متغیرهای توضیحی) تعریف نمی‌شوند. سیگل (۱۹۸۲) برآوردگر میانه مکرر[2] را با ضریب شکست ۵۰ درصد معرفی نمود، ولی ضعف این روش در عدم تشخیص بخش‌های «خوب» و «بد» نمونه بود. روسو (۱۹۸۴) برآوردگر کمترین میانه توان‌های دوم[3] را که براساس طرحی از همپل (۱۹۷۵) بود، به صورت $\min_\beta \{\mathrm{median}_{1 \leq i \leq n} \ e_i^2\}$ معرفی کرد. این روش با نسبت به تبدیلات خطی در متغیرهای توضیحی هموردا[4] (اندازه تغییرات هماهنگ دو متغیر تصادفی) است. عدم سادگی مینیمم‌سازی تابع هدف

[1] Contaminated
[2] Repeated median estimator
[3] Least median squares
[4] Equivariant

در این مورد منجر به معرفی برآوردگر کمترین‌توان‌های دوم پیراسته[1] توسط روسو و لروی (۱۹۸۷) شد که در ادامه مورد بررسی قرار می‌گیرد.

در بخش ۲ رگرسیون استوار توان‌های دوم پیراسته معرفی شده است. مروری بر رگرسیون ستیغی و معرفی آن در بخش ۳ آمده است. در بخش ۴ کاربردی از این روش برای داده‌های سن دندانی مورد مطالعه قرار گرفته است.

۳ رگرسیون استوار کمترین توان‌های دوم پیراسته

فرض کنید $(e^۲)_{۱:n} \leq (e^۲)_{۲:n} \leq (e^۲)_{i:n} \leq (e^۲)_{n:n}$ توان دوم مانده‌های مرتب شده هستند. لازم به ذکر است که ابتدا مانده‌ها به توان دو می‌رسند، سپس مرتب می‌شوند. در روش رگرسیون استوار کمترین توان‌های دوم پیراسته، مجموع h کوچکترین توان‌های دوم مانده‌های مدل به صورت $\sum_{i=۱}^{h}(e^۲)_{i:n}$ کمینه می‌شوند. h در بازه $[\frac{n}{۲}, n]$ تغییر می‌کند و $\alpha = \frac{n-h}{h}$ درصد مشاهدات دور افتاده را نشان می‌دهد. روسو و لروی (۱۹۸۷) نشان داد که با انتخاب $h = (\frac{n}{۲}) + (\frac{p+۱}{۲})$ نقطه شکست روش کمترین توان‌های دوم پیراسته به ماکسیمم مقدار ممکن خواهد رسید. اما برای تعیین اینکه مشاهده iام، یک مشاهده خوب است یا نه لازم است تابع نشانگر z_i را به صورت

$$z_i = \begin{cases} ۱ & \text{مشاهده } i\text{ام دورافتاده نباشد} \\ ۰ & \text{مشاهده } i\text{ام دورافتاده باشد} \end{cases}.$$

تعریف می‌شود. در نهایت ماتریس $\mathbf{Z}$ که یک ماتریس قطری است به صورت

$$\mathbf{Z} = \begin{pmatrix} z_۱ & \cdots & ۰ \\ \vdots & \ddots & \vdots \\ ۰ & \cdots & z_n \end{pmatrix},$$

به دست می‌آید. شکل ماتریسی مدل رگرسیون خطی به صورت

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = ۰, \quad E(\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}) = \sigma^۲ \boldsymbol{I}_p,$$

[1] Least trimmed squares

خواهد بود، به‌طوری‌که $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ بردار $n \times 1$ از خطاهای تصادفی است. بنابراین می‌توان مسئله کمینه‌سازی را به صورت

$$\begin{aligned} \min_{\boldsymbol{\beta}, \mathbf{Z}} \varphi(\boldsymbol{\beta}, \mathbf{Z}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{Z} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \mathbf{e}^\top \mathbf{z} &= h, \quad z_i \in \{0, 1\}, \quad i = 1, \ldots, n, \end{aligned} \tag{1}$$

بازنویسی کرد، که در آن $\mathbf{e} = (1, \ldots, 1)^\top_{n \times 1}$ و $\mathbf{z} = (z_1, \ldots, z_n)$. با حل مسئله بهینه‌سازی (۱)، برآوردگر کمترین توان‌های دوم پیراسته به صورت

$$\hat{\boldsymbol{\beta}}(\mathbf{Z}) = \mathbf{C}_{\mathbf{Z}}^{-1} \mathbf{X}^\top \mathbf{Z} \mathbf{y}, \quad \mathbf{C}_{\mathbf{z}} = \mathbf{X}^\top \mathbf{Z} \mathbf{X},$$

به‌دست می‌آید. در صد شکست این روش ۵۰ درصد است. یعنی حتی اگر نیمی از داده‌ها دورافتاده باشند باز هم این روش در برابر نقاط دورافتاده، استوار است (روزبه ، ۲۰۱۶؛ امینی و روزبه ، ۲۰۱۶؛ روزبه و آرشی ، ۲۰۱۷).

۴ رگرسیون ستیغی استوار

وجود ارتباط خطی بین متغیرهای توضیحی در مدل رگرسیون خطی چندگانه، همخطی نامیده می‌شود. دو نوع همخطی وجود دارد: همخطی کامل و همخطی ناقص. همخطی کامل، زمانی رخ می‌دهد که یکی از متغیرهای توضیحی، تابعی دقیق از یک یا چند متغیر توضیحی باشد و همخطی ناقص، زمانی رخ می‌دهد که یکی از متغیرهای توضیحی، تابعی تقریبی از یک یا چند متغیر توضیحی باشند. در عمل به دلیل وجود خطای اندازه‌گیری، همخطی کامل رخ نمی‌دهد. وجود همخطی ممکن است منجر به فاصله‌های اطمینان پهن برای پارامترها، ناپایداری برآوردها یا تولید برآوردهایی با علامت اشتباه شود. روش‌های مختلفی برای غلبه برای همخطی وجود دارد که از آن جمله رگرسیون مولفه اصلی و رگرسیون ستیغی اشاره کرد. هورل و کنارد (۱۹۷۰) برای رفع مشکل همخطی برآوردگر ستیغی را با کمینه‌سازی تابع هدف به صورت

$$\varphi(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad \boldsymbol{\beta}^\top \boldsymbol{\beta} \leq \phi^2, \tag{2}$$

معرفی کردند، که در آن ϕ[۲] پارامتر منظم‌سازی[۱] بوده و برای کنترل ضرایب برآورد شده بکار می‌رود. از کمینه‌سازی تابع هدف (۲) برآوردگر ستیغی به صورت

$$\hat{\boldsymbol{\beta}}(k) = \mathbf{C}_k^{-1}\mathbf{X}^\top\mathbf{y}, \quad \mathbf{C} = \mathbf{X}^\top\mathbf{X}, \quad \mathbf{C}_k = \mathbf{C} + k\mathbf{I}_p,$$

به‌دست می‌آید. این برآوردگر، یک برآوردگر اریب است و به k پارامتر ستیغی یا انقباضی گفته می‌شود. نکته مهم در روش ستیغی یافتن مقدار k است چرا که مقدار زیاد k باعث افزایش اریبی و کاهش واریانس و مقدار کم k باعث کاهش اریبی و افزایش واریانس می‌شود (امینی و روزبه ، ۲۰۱۵) و برای حداقل نمودن تابع مخاطره لازم است k به گونه‌ای تعیین شود که موازنه بین اریبی و واریانس برقرار کند. روش‌های مختلفی مانند استفاده از معیارهای آکائیک، اطلاع بیزی، C_p مالوز، اعتبارسنجی متقابل، اعتبارسنجی متقابل تعمیم‌یافته و مانند آن‌ها برای انتخاب مقدار k وجود دارد. برای اطلاعات بیشتر در مورد برآوردگر رگرسیونی ستیغی می‌توان به راسخ و همکاران (۱۳۹۸) و امامی (۱۳۹۷) مراجعه کرد. واسرمن (۲۰۰۶) از بین روش‌های بیان‌شده روش اعتبار سنجی و اعتبارسنجی متقابل را به دلیل یافتن همزمان مقدار بهینه k و آزمون میزان دقت برگزیدند.

مجموع مجذور خطا با روش اعتبارسنجی[۲] متقابل به صورت $\mathrm{CV} = \|\mathbf{y} - \hat{\mathbf{y}}_{(-i)}\|^2$ محاسبه می‌شود. که در آن $\hat{\mathbf{y}}_{(-i)} = \mathbf{X}_{(-i)}\hat{\boldsymbol{\beta}}_{(-i)}$، $\hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}_{(-i)}^\top\mathbf{X}_{(-i)})^{-1}\mathbf{X}_{(-i)}^\top\mathbf{y}_{(-i)}$ و $\mathbf{X}_{(-i)}$ از حذف iامین سطر ماتریس $\mathbf{X}$ و $\mathbf{y}_{(-i)}$ نیز از حذف iامین مشاهده بردار $\mathbf{y}$ بدست می‌آید. برای مشاهده جزئیات بیشتر به واسرمن (۲۰۰۶) مراجعه نمایید.

در روش اعتبارسنجی متقابل در هر مرحله باید یکی از مشاهدات حذف و مدل بدون آن مشاهده برازش شود یعنی در واقع باید n بار برازش انجام شود تا بتوان مقدار k بهینه را یافت و این امر باعث دشواری محاسبات و زمان‌گیر شدن آن خواهد شد. به همین دلیل روش اعتبارسنجی متقابل تعمیم‌یافته[۳] که محاسبات آن ساده‌تر از اعتبارسنجی متقابل بوده و به صورت

$$\mathrm{GCV}_k = \frac{\frac{1}{n}\|(\mathbf{I}_n - \mathbf{H}_k)\mathbf{y}\|^2}{(1 - \frac{1}{n}\mathrm{tr}(\mathbf{H}_k))^2},$$

تعریف می‌شود، پیشنهاد شد (واسرمن ، ۲۰۰۶)، که در آن $\mathbf{H}_k = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^\top$. در بسیاری

[1] Regularization parameter
[2] Cross-validation (CV)
[3] Generalized cross-validation (GCV)

از مسائل رگرسیونی ممکن است همزمان هم مشکل نقاط دورافتاده و هم مشکل همخطی وجود داشته باشد. ایده‌ای که در این‌گونه موارد به ذهن می‌رسد، استفاده همزمان از روش‌های ستیغی و استوار است، یعنی

$$\min_{\boldsymbol{\beta},\mathbf{Z}} \varphi(\boldsymbol{\beta},\mathbf{Z}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top \mathbf{Z}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \quad \boldsymbol{\beta}^\top\boldsymbol{\beta} \le \phi^2, \quad \mathbf{e}^\top\mathbf{z} = h, \quad z_i \in \{0,1\}. \quad (۳)$$

با حل مسئله بهینه‌سازی (۳)، برآوردگر ستیغی استوار به صورت $\hat{\boldsymbol{\beta}}(k,\mathbf{Z}) = \mathbf{C}_{k,\mathbf{Z}}^{-1}\mathbf{X}^\top\mathbf{Z}\mathbf{y}$ به‌دست می‌آید، که در آن $\mathbf{C}_{k,\mathbf{Z}} = \mathbf{C}_\mathbf{Z} + k\mathbf{I}_p$. برای محاسبه برآوردگر ستیغی استوار در مسائل عددی، ابتدا برآوردگر کمترین توان‌های دوم پیراسته را با استفاده از رابطه (۱) محاسبه نمود و ماتریس $\mathbf{Z}$ ساخته می‌شود. سپس، معیار اعتبارسنجی متقابل تعمیم‌یافته را برای برآوردگر ستیغی استوار به‌صورت

$$\mathrm{GCV}_{k,\mathbf{z}} = \frac{\frac{1}{n}\|(\mathbf{I}_n - \mathbf{H}_{k,\mathbf{z}})\mathbf{y}\|^2}{(1-\frac{1}{n}\mathrm{tr}(\mathbf{H}_{k,\mathbf{z}}))^2}, \quad \mathbf{H}_{k,\mathbf{z}} = \mathbf{X}\mathbf{C}_{k,\mathbf{z}}^{-1}\mathbf{X}^\top,$$

بازنویسی می‌شود. در انتها با کمینه‌سازی $\mathrm{GCV}_{k,\mathbf{z}}$ به کمک دستور optim در نرم‌افزار R مقدار بهینه k را به روش عددی L-BFGS-B بدست آورده و در رابطه (۳) جایگزین می‌شود.

۵ مطالعه شبیه‌سازی

در این بخش در یک مطالعه شبیه‌سازی به بررسی و مقایسه چهار برآوردگر کمترین توان‌های دوم، ستیغی، کمترین توان‌های دوم پیراسته و ستیغی استوار پرداخته می‌شود. در این مطالعه بردار پارامترها به‌صورت $\boldsymbol{\beta} = (-3, 4, 5.5, 2, -4, 7)^\top$ در نظر گرفته شده است. به منظور ایجاد همخطی بین متغیرهای توضیحی برای تولید هر یک از درایه‌های ماتریس طرح از رابطه

$$x_{ij} = (1-\gamma^2)^{\frac{1}{2}} z_{ij} + \gamma z_{ip}, \quad i = 1,\ldots,n, \quad j = 1,\ldots,p,$$

استفاده شده است. به‌طوری که در آن z_{ij} اعداد شبه‌تصادفی نرمال استاندارد مستقل بوده و γ طوری تعیین می‌شود که بین هر دو متغیر توضیحی همبستگی ایجاد شود. در این‌جا $\gamma = 0.75, 0.87$ فرض می‌شود. از طرفی برای ایجاد نقاط پرت در مجموعه داده‌ها، ۷۵ درصد خطاها از توزیع نرمال استاندارد و ۲۵ درصد

بقیه از توزیع $t-$ استیودنت غیرمرکزی تولید می‌شود، یعنی

$$\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_۱^\top, \boldsymbol{\varepsilon}_۲^\top)^\top \quad \boldsymbol{\varepsilon}_۱ \sim N(۰, \mathbf{I}), \quad \boldsymbol{\varepsilon}_۲ \sim t_۲(۸),$$

بطوریکه $t_\nu(\delta)$ نشان دهنده توزیع $t-$استیودنت غیر مرکزی با ν درجه آزادی و پارامتر غیرمرکزی δ است. در نهایت متغیر پاسخ با $n = ۱۰۰$ مشاهده و ۱۰۰۰ تکرار از مدل $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ تولید می‌شود. لازم به ذکر است که محاسبات در سیستمی با ویژگی‌های زیر انجام شده است:

Intel(R) Pentium(R) CPU G3250 @ 3.20 GHz 3.20 GHz, 8 GB RAM,
System type 64-bit Operating System, x64-based processor.

مقادیر برآورد پارامترها، مقادیر اریبی، انحراف استاندارد و همچنین $\text{SSE} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$ برآوردگرهای پیشنهادی در جدول‌های ۱، ۲، ۳، ۴ و ۵ گزارش شده است. با توجه به این جدول‌ها کاملا واضح است که روش ستیغی استوار بسیار بهتر از روش‌های دیگر عمل می‌کند، چرا که کمترین میزان SSE متعلق به این روش بوده است. همچنین میزان اریبی و انحراف استاندارد براوردگر ستیغی استوار در مقایسه با روش‌های دیگر کمتر است.

جدول ۱. ارزیابی برآوردگرهای کمترین توان‌های دوم معمولی و پیراسته برای $\gamma = ۰/۷۵$.

روش	معمولی			پیراسته		
	برآورد	اریبی	انحراف استاندارد	برآورد	اریبی	انحراف استاندارد
$\beta_۱$	−۲/۶۹۶۶	۰/۳۰۳۳	۲/۳۷۱۴	−۲/۸۸۲۲	۰/۱۱۷۷	۲/۰۷۸۳
$\beta_۲$	۴/۱۷۷۶	۰/۱۷۷۶	۱/۷۸۶۰	۳۹۵۱۴	−۰/۰۴۸۵	۱/۴۸۰۸
$\beta_۳$	۵/۵۶۵۸	۰/۰۶۵۸	۱/۴۹۸۴	۵۵۴۷۶	۰/۰۴۷۶	۱/۲۵۸۸
$\beta_۴$	۲/۳۷۷۳	۰/۳۷۷۳	۲/۲۰۹۷	۱۸۶۰۲	−۰/۱۳۹۷	۱/۷۶۶۷
$\beta_۵$	−۳/۸۴۷۹	۰/۱۵۲۰	۲/۹۹۲۴	−۳۹۴۲۲	۰/۰۵۷۷	۲/۵۳۴۸
$\beta_۶$	۶/۴۰۱۳	−۰/۵۹۸۶	۳/۴۰۱۷	۶۹۷۹۸	−۰/۰۲۰۱	۲/۵۴۹۳

جدول ۲. ارزیابی برآوردگرهای ستیغی و ستیغی استوار برای $\gamma = ۰/۷۵$.

روش	ستیغی			ستیغی استوار		
	برآورد	اریبی	انحراف استاندارد	برآورد	اریبی	انحراف استاندارد
$\beta_۱$	−۲/۱۸۰۸	۰/۸۱۹۱	۰/۶۵۴۶	−۲/۸۰۶۳	۰/۱۹۳۶	۰/۶۴۹۷
$\beta_۲$	۴/۰۰۵۰	۰/۰۰۵۰	۰/۵۸۵۴	۳۹۴۵۱	−۰/۰۵۴۸	۰/۵۷۴۲
$\beta_۳$	۵/۲۴۲۰	−۰/۲۵۷۹	۰/۵۸۵۸	۵۵۲۱۹	۰/۰۲۱۹	۰/۵۷۸۹
$\beta_۴$	۲/۴۰۴۹	۰/۴۰۴۹	۰/۷۵۲۲	۱۸۸۰۹	−۰/۱۱۹۰	۰/۷۴۰۸
$\beta_۵$	−۳/۱۹۸۴	۰/۸۰۱۵	۰/۶۲۹۷	−۳۸۵۲۲	۰/۱۴۷۷	۰/۶۳۱۰
$\beta_۶$	۵/۹۰۴۲	−۱/۰۹۵۷	۰/۸۷۵۹	۶۸۸۱۲	−۰/۱۱۸۷	۰/۸۶۳۸

جدول ۳. ارزیابی برآوردگرهای کمترین توان‌های دوم معمولی و پیراسته برای $\gamma = ۰/۸۷$.

روش	کمترین توان دوم معمولی			کمترین توان دوم پیراسته		
	برآورد	اریبی	انحراف استاندارد	برآورد	اریبی	انحراف استاندارد
$\beta_۱$	−۲٫۸۳۲۸	۰٫۱۶۷۱	۲٫۱۵۸۷	−۲٫۹۷۰۴	۰٫۰۲۹۵	۰٫۷۷۳۲
$\beta_۲$	۳٫۹۸۱۰	−۰٫۰۱۸۹	۲٫۰۶۶۵	۳٫۹۱۴۹	−۰٫۰۸۵۰	۰٫۸۴۳۷
$\beta_۳$	۵٫۶۵۳۹	۰٫۱۵۳۹	۲٫۴۲۲۶	۵٫۵۶۳۶	۰٫۰۶۳۶	۰٫۸۱۵۱
$\beta_۴$	۱٫۹۷۳۲	−۰٫۰۲۶۷	۳٫۴۶۸۸	۱٫۹۹۴۵	−۰٫۰۰۵۴	۰٫۸۱۶۷
$\beta_۵$	−۳٫۸۸۷۶	۰٫۱۱۲۳	۲٫۰۶۸۵	−۴٫۱۳۶۲	−۰٫۱۳۶۲	۰٫۸۵۶۸
$\beta_۶$	۷٫۰۳۷۳	۰٫۰۳۷۳	۳٫۷۴۷۷	۷٫۰۷۶۶	۰٫۰۷۶۶	۱٫۱۹۳۷

جدول ۴. ارزیابی برآوردگرهای ستیغی و ستیغی استوار برای $\gamma = ۰/۸۷$.

روش	ستیغی			ستیغی استوار		
	برآورد	اریبی	انحراف استاندارد	برآورد	اریبی	انحراف استاندارد
$\hat{\beta}_۱$	−۱٫۸۵۵۹	۱٫۱۴۴۰	۱٫۷۰۷۱	−۲٫۸۳۸۲	۰٫۱۶۱۷	۰٫۷۷۶۲
$\hat{\beta}_۲$	۳٫۹۰۰۰	−۰٫۰۹۹۹	۱٫۵۹۱۹	۳٫۹۱۶۹	−۰٫۰۸۳۰	۰٫۸۲۰۵
$\hat{\beta}_۳$	۵٫۲۷۸۱	−۰٫۲۲۱۸	۱٫۹۲۰۳	۵٫۵۳۳۵	۰٫۰۳۳۵	۰٫۷۹۱۵
$\hat{\beta}_۴$	۲٫۱۶۸۰	۰٫۱۶۸۰	۲٫۴۸۴۸	۲٫۰۳۲۴	۰٫۰۳۲۴	۰٫۷۹۷۴
$\hat{\beta}_۵$	−۲٫۶۹۱۰	۱٫۳۰۸۹	۱٫۸۰۶۱	−۳٫۹۷۶۵	۰٫۰۲۳۴	۰٫۸۵۵۳
$\hat{\beta}_۶$	۵٫۶۷۷۱	−۱٫۳۲۲۸	۲٫۳۰۵۴	۶٫۸۶۷۶	−۰٫۱۳۲۴	۱٫۱۸۱۶

جدول ۵. خطا و سرعت محاسبه برآوردگرهای پیشنهادی.

روش برآورد	$\gamma = ۰/۷۵$		$\gamma = ۰/۸۷$	
	SSE	سرعت محاسبات	SSE	سرعت محاسبات
کمترین توان‌های دوم	۳۶٫۷۵۵۸۰۲	۰۰ : ۰۰ : ۰۶	۴۴٫۷۷۲۷۴۰	۰۰ : ۰۰ : ۰۹
کمترین توان‌های دوم پیراسته	۲٫۸۵۷۴۸۸	۰۰ : ۰۵ : ۵۴	۴٫۷۸۹۳۰۴	۰۰ : ۰۶ : ۰۷
ستیغی	۲۶٫۶۴۵۹۵۱	۰۰ : ۰۱ : ۵۱	۲۸٫۵۰۸۷۰۲	۰۰ : ۰۱ : ۵۷
ستیغی استوار	۲٫۸۴۳۸۰۰	۰۰ : ۰۶ : ۴۶	۴٫۶۷۲۹۵۵	۰۰ : ۰۶ : ۵۶

۶ مطالعه کاربردی: تحلیل داده‌های سن دندانی

برآورد سن، نقش مهمی در پزشکی قانونی، بیماری‌های غدد، دندان پزشکی، باستان شناسی و بسیاری علوم دیگر دارد. در کشورهای در حال توسعه که ثبت دقیق زمان تولد افراد وجود ندارد، این روش راه بسیار مناسبی برای تخمین سن تقویمی افراد است. اغلب بین بلوغ فیزیکی و سن فرد ارتباط وجود دارد و برخی از نشانه‌های بلوغ فیزیکی مانند سن اسکلتی، قاعدگی، اندازه و قد افراد و کلسیفیکاسیون (جمع شدن کلسیم در یکی از بافت‌های بدن) می‌تواند در نبود اطلاعاتی نظیر سن و جنسیت دقیق افراد، برای تعیین سن افراد مورد استفاده قرارگیرد. سن دندانی می‌تواند بر اساس رویش یا مراحل شکل‌گیری دندان در رادیوگرافی ارزیابی شود. در این بین مدل‌سازی برای برآورد سن تقویمی با استفاده از روش‌های نوین و غنی آماری دارای اهمیت است. در بین روش‌هایی که برای تعیین سن استفاده می‌شود، رایج‌ترین روش

مورد استفاده در سراسر دنیا، روش دمیرجیان است. این روش بر اساس کلسیفیکاسیون دندان دائمی در رادیوگرافی پانورامیک است. در این نوع رادیوگرافی یک نوع اسکن پانورامیک فک بالا و پایین به وسیله اشعه ایکس دندان پزشکی است که دیدی دو بعدی از نیمکره فرضی تشکیل‌شده در حد فاصل بین گوش‌ها می‌دهد، است. مشکلی که در بررسی سن دندانی با روش دمیرجیان وجود دارد زمانی است که تقریبا تمام ریشه‌های دندان‌های دائمی بیمار تکامل‌یافته و ریشه‌های دندان‌های عقل بیمار باید به عنوان ملاک مورد بررسی قرار گیرند. دمیرجیان و همکارانش روش جدیدی ابداع کردند (دمیرجیان و همکاران ، ۱۹۷۶۳). پس از رویش دندان مولر دوم دائمی در حدود سنین ۱۲ تا ۱۳ سالگی تعیین سن با مشکل مواجه می‌شود. این مسئله، دندان مولر سوم را در این سنین در مرکز توجه قرار می‌دهد چرا که هنوز در حال شکل‌گیری است. این دندان در سنین سالگی در دهان ظاهر می‌شود و با توجه به فرآیند تشکیل ریشه می‌توان مراحل تکاملی آن ۱۷ تا ۲۱ را تعریف کرد که تقریبا از ۱۴ سالگی آغاز می‌شود. درکنار استفاده از روش دمیرجیان در تعیین سن با کمک دندان مولر سوم، روش‌های دیگری نیز وجود دارد که برپایه تحقیقات صورت گرفته روش دمیرجیان بر سایر روش‌ها عملکرد بهتری داشته است. اردکانی و همکاران (۲۰۰۷) مطالعه‌ای با هدف مطالعه مقطعی از نوع تشخیصی با همکاری سازمان پزشکی قانونی و دانشکده دندانپزشکی شهید صدوقی یزد در سال ۸۴ – ۱۳۸۳ درباره ۵۸ بیمار با محدوده سنی ۱۵ تا ۲۵ سال که به منظور جراحی دندان عقل به مطب خصوصی مراجعه کرده بودند، صورت گرفت. محاسبات آماری از طریق آزمون‌های تی زوجی، ضریب همبستگی، ویلکاکسون و کلموگروف اسمیرینف انجام پذیرفت. کمترین خطاهای به دست آمده در تخمین سن با سن تقویمی بیمار، مربوط به دندان عقل کشیده شده سمت راست فک پایین بود.

در این مقاله، بیماران مراجعه کننده در طی دوره ۱۲ماهه سال ۹۴ – ۱۳۹۳ به درمانگاه خاتم‌الانبیا استان یزد، جامعه مورد بررسی تحقیق را تشکیل داده است. تعداد ۸۷ نفر از بیماران مراجعه کننده، که به صورت روش تصادفی ساده انتخاب شده اند. معیارهای ورود و خروج نمونه‌ها در مطالعه عبارتند از عدم وجود مشکلات رشدی- تکاملی، کیفیت قابل قبول رادیوگرافی پانورامیک، وجود حداقل یک دندان عقل و عدم وجود شکستگی و یا ضایعاتی پیرامون دندان عقل. تعداد ۸۷ رادیوگرافی پانورامیک دیجیتال مربوط به بیماران نوجوان ارتودنسی مراجعه‌کننده به کلینیک خاتم‌الانبیا از آرشیو کلینیک بررسی شد. جنسیت و سن کرونولوژیک دقیق افراد (دارای روز و ماه تولد علاوه برسال) از روی پرونده بیماران ثبت شد، در صورت عدم وجود در پرونده از طریق تماس با بیمار به طور مستقیم پرسیده شد. رادیوگرافی‌ها دارای دندان‌های عقل بوده و بین سنین ۱۴ تا ۱۹ قرار داشت.

در جدول‌های ۶ و ۷ به ترتیب درصد فراوانی بیماران به تفکیک جنسیت و سن تقویمی و برآوردشده

جدول ۶. فراوانی داده‌ها به تفکیک جنسیت.

جنسیت	زن	مرد
فراوانی نسبی	۶۷٫۹	۳۲٫۱

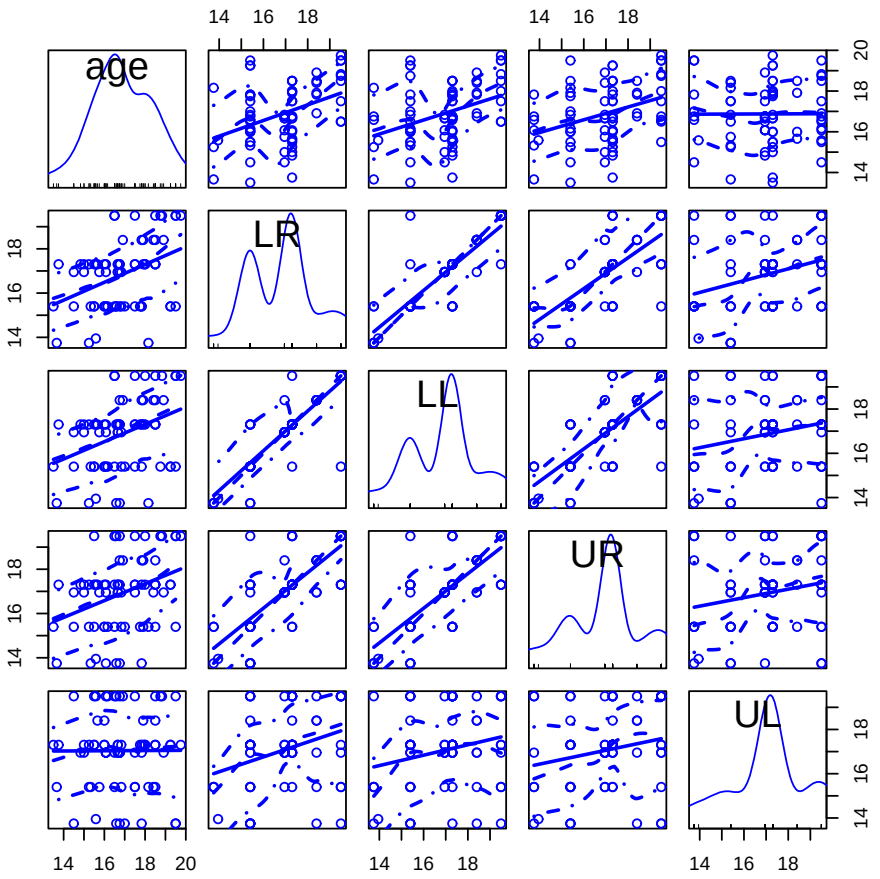
به نمایش گذاشته شده است. مدل رگرسیون خطی چندگانه به صورت

جدول ۷. آماره‌های توصیفی به تفکیک سن تقویمی و آزمون تخمین دمیرجیان.

–	مینیمم	ماکسیمم	میانگین	انحراف معیار
سن تقویمی	۱۳٫۵	۲۰٫۵۸	۱۷٫۲۱	۱٫۱۶۶
آزمون دمیرجیان	۱۳٫۷۵	۱۹٫۵	۱۶٫۸۸	۱٫۰۸

$$(age)_i = \beta_1(LR)_i + \beta_2(LL)_i + \beta_3(UR)_i + \beta_4(UL)_i + \varepsilon_i \quad i = 1, \ldots, 70, \qquad (۴)$$

نوشته می‌شود، که در آن age متغیر پاسخ سن تقویمی، LR سن برآوردشده دمیرجیان دندان راست مولر فک پایین، LL سن برآوردشده دمیرجیان دندان چپ مولر فک پایین، UR سن برآوردشده دمیرجیان دندان راست مولر فک بالا و UL سن برآوردشده دمیرجیان دندان چپ مولر فک بالا هستند. همان‌طور که در شکل ۱ ملاحظه می‌شود، متغیرهای توضیحی دارای رابطه خطی معنی‌داری با متغیر وابسته نیستند. با انجام تبدیلات باکس – کاکس روی داده‌ها، تا حد ممکن رابطه خطی معنی دارتری بین متغیرهای توضیحی و وابسته ایجاد می‌شود، بطوریکه دست‌یابی به یک مدل رگرسیون قوی‌تر را امکان‌پذیر می‌کند. تبدیل باکس کاکس روی داده‌های اکیدا مثبت انجام می‌شود. در واقع این تبدیل توزیع متغیرها را به نرمال چند متغیره نزدیک می‌کند (کاکس و باکس ، ۱۹۶۴). با انجام این تبدیل که به صورت نمودارهای باکس – کاکس در شکل ۲ به نمایش گذاشته شده‌است، وجود رابطه خطی معنی‌دارتری بین متغیرهای توضیحی و وابسته نسبت به حالت قبل مشهود است. برای انجام این تبدیلات از نرم‌افزار مینی‌تب استفاده شد. برای بررسی نحوه ارتباط خطی بین متغیرهای توضیحی از نمودار ۳ با استفاده از بسته "PerformanceAnalytics" در نرم‌افزار R 3.4.4 رسم شده است. همان‌طور که ملاحظه می‌شود، بیشترین مقدار همبستگی ۰٫۸۷ است که مربوط به دو متغیر توضیحی $LRnew$ و $LLnew$ برای بررسی وجود همخطی در مدل از آماره کاپا

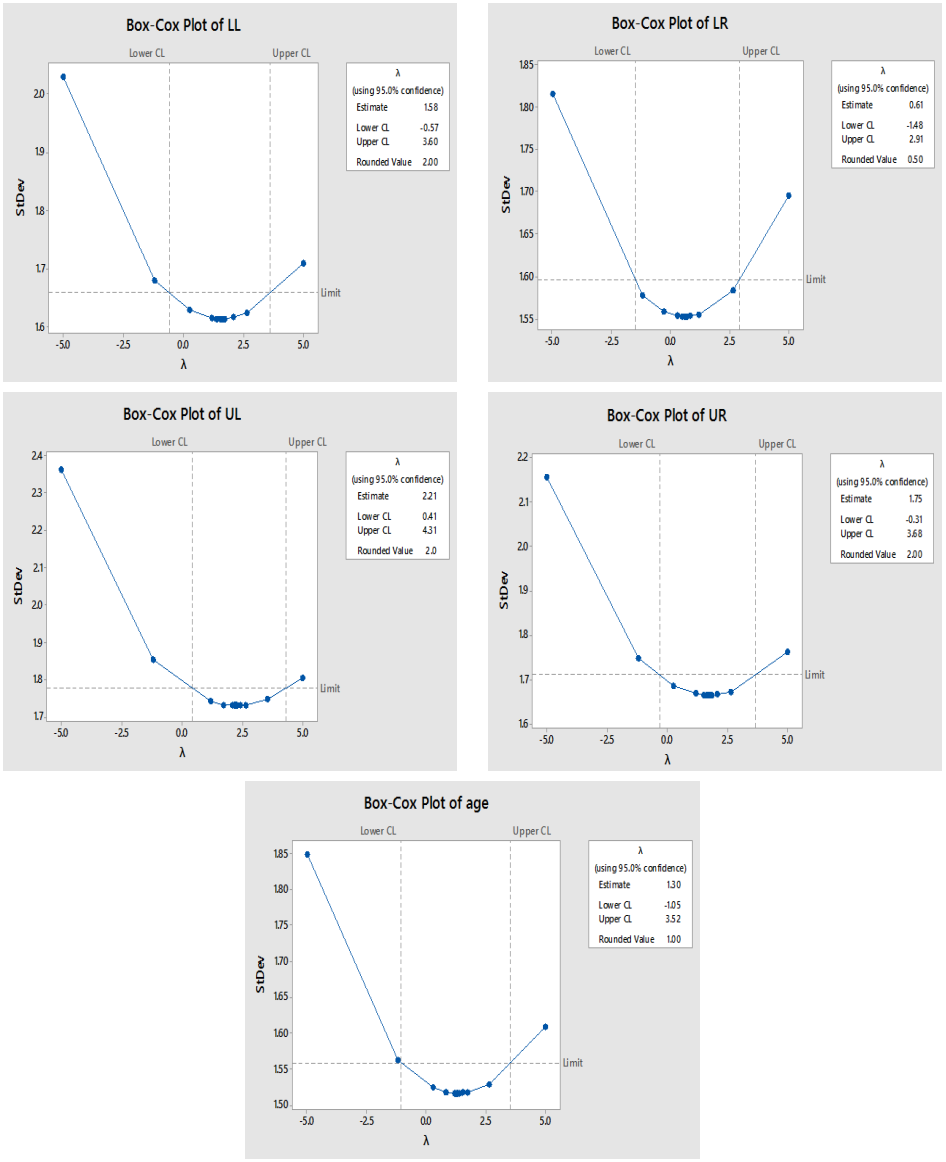


شکل ۱. نمودار پراکنش متغیرهای پاسخ و توضیحی.

به صورت

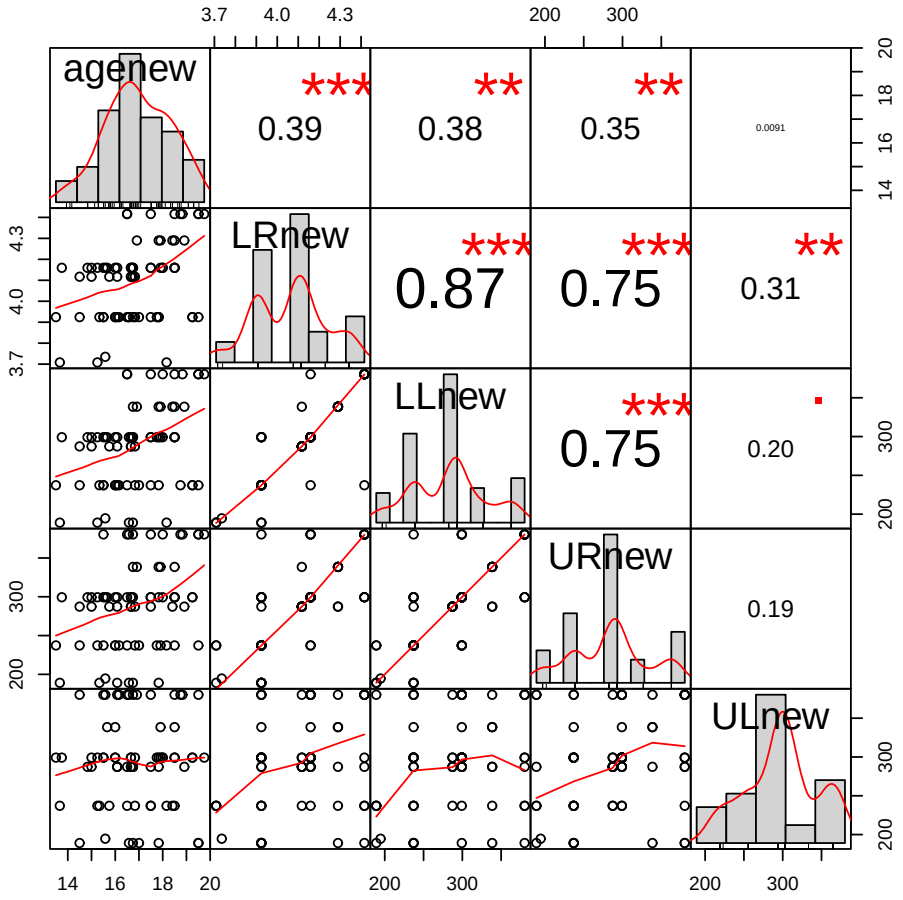
$$\text{kappa} = \sqrt{\frac{\max \lambda_j}{\min \lambda_j}} = ۱۴۶٫۹۲۱۵, \quad j = ۱, \ldots, p,$$

استفاده می‌شود، که در آن λ_i ها مقادیر ویژه ماتریس $\mathbf{X}^\top \mathbf{X}$ در دادهای تبدیل یافته است. چون مقدار آماره کاپا بیشتر از ۱۰۰ است، شکل همخطی جدی وجود دارد.



شکل ۲. تبدیلات باکس – کاکس بر روی متغیرهای توضیحی.

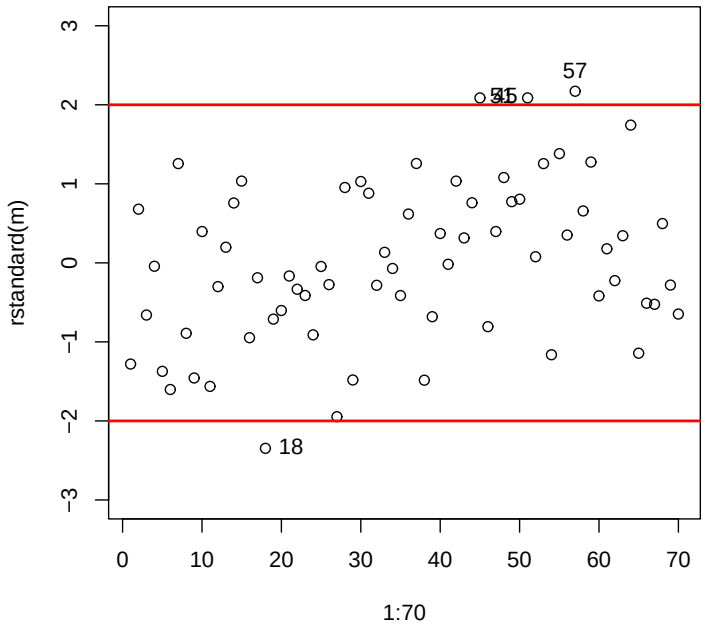
در ادامه، وجود نقاط دورافتاده در مدل مورد بررسی قرار می‌گیرد. اگر r_i مانده استانداردشده مشاهده iام را نشان دهد، به‌ازای $|r_i| > ۲$ باشد، مشاهده دورافتاده است. (شیدر ، ۲۰۰۹). برای تشخیص



شکل ۳. نمودار همبستگی متغیرهای تبدیل‌یافته.

نقاط دورافتاده در مدل از شکل ۴ استفاده می‌شود. همان‌طور که لاحظه می‌شود چهار مشاهده ۱۸ام، ۴۵ام، ۵۱ام و ۵۷ام دورافتاده‌اند. بنابراین به‌طور همزمان مشکل همخطی و نقاط دورافتاده در مدل رگرسیونی (۲) مشاهده می‌شود. درادامه برای مقابله با این مشکلات از رگرسیون ستیغی استوار استفاده می‌کنیم. در این‌جا پارامترها به چهار روش کمترین توان‌های دوم، روش ستیغی، روش کمترین توان‌های دوم پیراسته، روش ستیغی کمترین توان‌های دوم پیراسته برآورد شده است و درنهایت این چهار روش را برای مدل نهایی

$$(agenew)_i = \beta_1(LRnew)_i + \beta_2(LLnew)_i + \beta_3(URnew)_i + \beta_4(ULnew)_i + \varepsilon_i$$



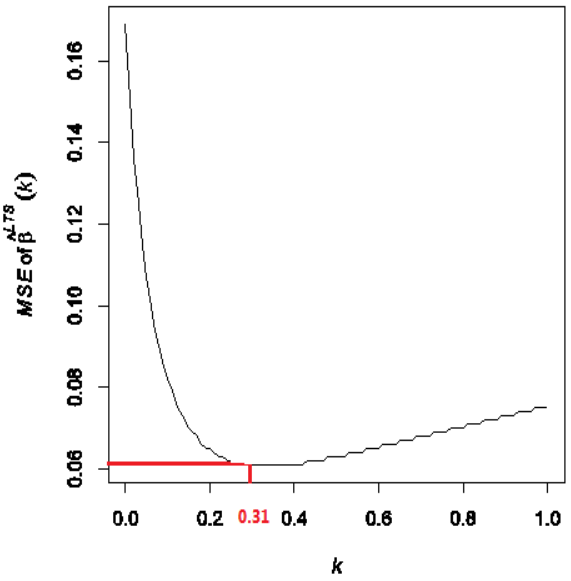
شکل ۴. نمودار مانده‌های استانداردشده.

$$i = ۱, \ldots, ۷۰, \qquad (۵)$$

با هم مقایسه می‌شوند، که در آن منظور از $agenew$، $LRnew$، $ULnew$، $URnew$ و $LLnew$ متغیرهای تبدیل‌یافته برپایه تبدیل باکس-کاکس شکل ۲ است. برای تعیین بهترین k از معیار اعتبارسنجی متقابل تعمیم‌یافته استفاده شده است.

جدول ۸. برآوردگرهای پیشنهادی به همراه انحراف استاندارد آن‌ها.

	کمترین توان دوم		کمترین توان دوم پیراسته		ستیغی		ستیغی استوار	
پارامتر	برآورد	SD	برآورد	SD	برآورد	SD	برآورد	SD
$\hat{\beta}_۱$	۰٫۲۸۴۸	۰٫۰۹۳	۰٫۳۸۴۵	۰٫۰۹۷۴	−۰٫۱۹۱۵	۰٫۰۴۶۴	−۰٫۲۵۳۹	۰٫۰۵۹
$\hat{\beta}_۲$	۰٫۰۷۲۰	۰٫۰۰۶	۰٫۱۷۴۰	۰٫۰۰۷	۰٫۰۹۳۹	۰٫۰۰۳	۰٫۰۹۶۵	۰٫۰۰۴
$\hat{\beta}_۳$	۰٫۱۰۴۶	۰٫۰۰۳	−۰٫۲۹۸	۰٫۰۰۳	−۰٫۰۲۰۰	۰٫۰۰۲	−۰٫۱۲۰۳	۰٫۰۰۱۹
$\hat{\beta}_۴$	۰٫۱۱۳۴	۰٫۰۰۰۴	−۰٫۰۸۶۱	۰٫۰۰۰۴	−۰٫۰۲۸۸	۰٫۰۰۰۲	−۰٫۰۴۷۱	۰٫۰۰۰۲



شکل ۵. نمودار مقادیر k در مقابل MSE برای مدل رگرسیونی ستیغی استوار.

جدول ۹. MSE و سرعت محاسبه چهار برآوردگر داده‌های سن تقویمی.

روش برآورد	MSE	سرعت محاسبات
کمترین توان‌های دوم	۱٫۱۷۰۹۱	۰۰ : ۰۰ : ۰۱
کمترین توان‌های دوم پیراسته	۰٫۱۶۷۳۱	۰۰ : ۰۰ : ۰۶
ستیغی	۰٫۰۸۱۰۲	۰۰ : ۰۰ : ۰۶
ستیغی استوار	۰٫۰۶۳۰۰	۰۰ : ۰۰ : ۰۹

مقدار بهینه برای مدل رگرسیونی ستیغی استوار برابر ۰٫۳۱ است که این موضوع در شکل ۵ به خوبی نشان داده شده است. همان‌طور که در جدول ۹ دیده می‌شود، کمترین میزان MSE مربوط به روش ستیغی استوار است، که نشان‌دهنده عملکرد بهتر این روش برای این مجموعه داده‌ها است.

بحث و نتیجه‌گیری

با توجه به کاربرد گسترده مدل‌های رگرسیونی در علوم پزشکی، بررسی دقیق و عمیق داده‌ها و متغیرها از اهمیت بسیار برخوردار است. زیرا عدم بررسی دقیق داده‌ها منجر به تحلیل و مدل‌بندی نادقیق رگرسیونی

می‌شود. وجود داده‌های دور افتاده، وجود همبستگی بین متغیرهای توضیحی و همچنین فرضیات زیربنایی مدل رگرسیونی از مواردی است که همواره باید قبل از انجام رگرسیون باید به آن توجه داشت. در این مقاله وجود همزمان داده دورافتاده و همخطی بین متغیرهای توضیحی مورد بررسی قرار گفت و روش رگرسیون ستیغی کمترین توان‌های دوم پیراسته برای مواجهه با آن پیشنهاد شد. به عنوان کاربردی از این روش، مطالعه‌ای با هدف برآورد سن براساس دندان‌های مولر مورد بررسی قرار گرفت. نتایج به‌دست آمده نشان می‌دهد که روش پیشنهادی در مقایسه با سایر روش‌ها MSE کمتری دارد. همچنین روش‌های ذکر شده در مقاله در داده‌های شبیه‌سازی شده نیز به کار برده شد. نتایج حاصل نشان داد که روش ستیغی استوار در شرایط وجود همخطی و نقاط پرت به طور همزمان بهتر از سایر برآوردگرها عمل می‌کند.

تقدیر و تشکر

نویسندگان مقاله ضمن تشکر از اعضای مجترم هیئت تحریریه مجله، از پیشنهادها و نظرات ارزشمند داوران و ویراستار محترم مقاله که موجب ارتقاء سطح آن گردید کمال تشکر و قدردانی را دارند.

مراجع

امامی، ه. (۱۳۹۷). روش‌های تشخیصی در مدل‌های خطی ریج نیمه‌پارامتری با خطا در اندازه‌گیری، *مجله مدل‌سازی پیشرفته ریاضی*، ۸، ۱۹–۴۸.

راسخ، ع.، منصوری، ب. و نرگس هدایت‌پور، ن. (۱۳۹۸)، شناسایی مشاهدات پرت در مدل رگرسیونی ریج تحت محدودیت‌های خطی تصادفی، *مجله علوم آماری ایران*، **۱۳**، ۱۱۷–۱۳۷.

Amini, M. and Roozbeh, M. (2015), Optimal Partial Ridge Estimation in Restricted Semiparametric Regression Models, *Journal of Multivariate Analysis*, **136**, 26–40.

Amini, M. and Roozbeh, M. (2016), Least Trimmed Squares Ridge Estimation in Partially Linear Regression Model, *Journal of Statistical Computation and Simulation*, **86**, 2780–276.

Andrews, D. F. (1974), A Robust Method for Multiple Linear Regression, *Technometrics*, 16, 523-531.

Bartlett, M. S. (1949), Fitting a Straight Line When Both Variables are Subject to Error, *Biometrics*, **5**, 207-212.

Bickel, P. J. (1973), On Some Analogues to Linear Combination of Order Statistics in Linear Model, *The Annals of Statistics*, **1**, 957-616.

Box, G.E., and Cox, D.R. (1964), An Analysis of Transformed Data, *Journal of Royal Statistical Society, Ser. B*, **39**, 211-252.

Brown, G. W. and Mood, A. M. (1951), *On Median Tests for Linear Hypotheses*, Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability, edited by J. Neyman, University of California Press, Berkeley and Los Angeles, 159-166.

Demirjian, A. and Goldstein, H. (1976), New Systems for Dental Maturity Based on Seven and Four Teeth. *Annals Of Human Biology*, **3(5)**, 411-21.

Donho, D. L. and Huber, P. J. (1983), *The Notion of Breakdown Point, in A Festschrift for Erich Lehmann*, edited by P. Bickel, K. Doksum, and J. L. Hodges, Jr., Wadsworth, Belmont, CA.

Edgeworth, F. Y. (1887), On Observation Relating to Several Quantities, *Hermathena*, **6**, 279-28.

Ezoddini Ardakani, F., Navab, A., A., Bashardoost, N., Mansoorian, H., Ahmadieh, M. H. and Sadat Correlation between Chronological, Skeletal, and Dental Age on Panoramic Radiography in Patients Referred to Yazd Dental Clinics on 2004-05. *Journal Dental School Shahid Beheshti University of Medical Sciences.*, **24(4)**, 474-484.

Hampel, F. R. (1971), A General Qualitative Definition of Robustness, *Annals of Mathematical Statistics*, **42**, 1887-1896.

Hampel, F. R. (1975), Beyond Location Parameters: Robust Concepts and Methods, *Bulletin International Statistical institute*, **46**, 375-382.

Hodges, J. L. (1974), *Efficiency in normal Samples and Tolerance of Extreme Values for Some Estimates of Location*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 163-168.

Hoerl A. E. and Kennard R. W. (1970), Ridge Regression: Biased Estimation for Non-orthogonal Problems, *Thechnometrics*, **12**, 55–67.

Jaeckel, L.A. (1972), Estimating Regression Coefficients by Minimizing the Dispersion of Residuals, *Annals of Mathematical Statistics*, **5**, 1449-1458.

Jureckova, J. (1971), Nonparametric Estimate of Regression Coefficients, *Annals of Mathematical Statistics*, **42**, 1328-1338.

Koenker, R. and Bassett, G.J. (1978), Regression Quantiles, *Econornetrica*, **46**, 33-50.

Maronna R. A., Martin R. A. and Yohai V. J. (2006), *Robust Statistics: Theory and Methods*, John Wiley and Sons, New York.

Nair, K. R. and Shrivastava, M. P. (1942), On a Simple Method of Curve Fitting, *Sankhya*, **6**, 121-132.

Rousseeuw, P.J (1984), Least Median of Squares Regression, *Journal of the American Statistical Association*, **79**, 871-880.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley, New York.

Roozbeh M. (2016), Robust Ridge Estimator in Restricted Semiparametric Regression Models, *Journal of Multivariate Analysis* **147**, 127-144.

Roozbeh M. and Arashi, M. (2017), Least-Trimmed Squares: Asymptotic Normality of Robust Estimator in Semiparametric Regression Models, *Journal of Statistical Computation and Simulation* **147**, 1147-1130.

Siegel, A.F. (1982), Robust Regression Using Repeated Medians, *Biometrika*, **69**, 242-244.

Sheather, S. J. (2009), *A Modern Approach to Regression with R*, Springer, College Station, Texas.

Tukey, J. W. (1970), *Exploratory Data Analysis (Limited Preliminary Edition)*, Addison-Wesley, Reading, MA.

Velleman, P. F. and Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury-Press, Boston.

Wald, A. (1940), The Fitting of Straight Lines if Both Variables are Subject to Error, *Annals of Mathematical Statistics*, **11**, 284-300.

Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer-Verlag, New York.

Journal of Statistical Sciences, Autumn and Winter, 2020
Vol. 14, No. 2, pp 409-428
DOI: 10.29252/jss.13.2.409

Modeling of Chronological Age Using Least Trimmed Squares Ridge Regression

Roozbeh, M., Manavi, M.
Departmet of statistics, Semnan University, Semnan, Iran.

Abstract: The popular method to estimation the parameters of a linear regression model is the ordinary least square method which, despite the simplicity of calculating and providing the BLUE estimator of parameters, in some situations leads to misleading solutions. For example, we can mention the problems of multi-collinearity and outliers in the data set. The least trimmed squares method which is one of the most popular of robust regression methods decreases the influence of outliers as much as possible. The main goal of this paper is to provide a robust ridge estimation in order to model dental age data. Among the methods used to determine age, the most popular method throughout the world is the modern modified Demirjian method that is based on the calcification of the permanent tooth in panoramic radiography. It has been shown that using the robust ridge estimator is leading to reduce the mean squared error in comparison with the OLS method. Also, the proposed estimators were evaluated in simulated data sets.

Keywords: Demirjian dental age, Least Trimmed Squares, Multicollinearity, Outliers, Ridge estimator.

Mathematics Subject Classification (2010): 62J07.