مجله علوم آماری، بهار و تابستان ۱۴۰۰
جلد ۱۵، شماره ۱، ص ۳۹ – ۶۰
DOI: 10.29252/jss.15.1.57
مقاله پژوهشی

کاربرد عملگرهای وزنی در مدل رگرسیون قدرمطلق انحرافات مرتب شده

جلال چاچی[۱]، علیرضا چاجی[۲]

[۱] گروه آمار، دانشکده علوم ریاضی و کامپیوتر، دانشگاه شهید چمران اهواز
[۲] گروه برق، دانشکده فنی مهندسی، دانشگاه صنعتی شهدای هویزه

تاریخ دریافت: ۱۳۹۸/۱۱/۰۹ تاریخ پذیرش و انتشار: ۱۳۹۹/۱۰/۱۹

چکیده:

در این مقاله رویکرد جدیدی در برآورد پارامترهای مدل رگرسیون خطی کمترین قدرمطلق انحرافات معرفی می‌شود که مبتنی بر مسائل بهینه‌سازی بر مبنای الحاق وزنی قدرمطلق انحرافات مرتب شده است. الحاق وزنی قدرمطلق انحرافات برازش مرتب شده در مسئله بهینه‌سازی در حالی که توابع نیکویی برازش مختلفی را بطور همزمان در مسئله مدل‌سازی در نظر می‌گیرد، توانایی تحلیل داده‌ها به منظور شناسایی نقاط دورافتاده را نیز فراهم می‌کند. بر این اساس این رویکرد تحت تاثیر مشاهدات دورافتاده قرار نمی‌گیرد و در هر مسئله متناسب با تعداد مشاهداتی که پتانسیل دورافتاده بودن را دارا هستند، به انتخاب بهترین برآوردگر مدل با بهینه‌ترین مقدار نقطه شکست در بین مجموعه‌ای از برآوردگرهای کاندید دیگر می‌پردازد. نیکویی برازش رویکرد پیشنهادی در مدل‌سازی داده‌های شبیه‌سازی شده و داده‌های واقعی در مهندسی آب با حضور مشاهدات دورافتاده تحلیل شده است. همچنین در انتها به تحلیل حساسیت برآوردگرها شامل بررسی معیارهای نااریبی و کارایی برآوردگرها پرداخته شده است.

واژه‌های کلیدی: رگرسیون وزنی، قدرمطلق انحرافات مرتب شده، نقطه شکست، توان‌های دوم خطا.

آدرس الکترونیکی نویسنده مسئول مقاله: جلال چاچی، jalal.chachi@scu.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J05, 62J02, 62G08.

۱ مقدمه

تحلیل و برازش مدل‌های رگرسیونی (خطی و غیرخطی، استوار و غیر استوار) یکی از پرکاربردترین و متداول‌ترین موضوعات در مدل‌سازی داده‌ها است یکی از شیوه‌های عمده در تحلیل اینگونه مدل‌ها رسم نمودار پراکنش مقادیر متغیرهای تبیینی در مقابل مقادیر متغیر وابسته است که میزان اهمیت متغیرهای تبیینی در تبیین متغیر وابسته را آشکار می‌سازد (آتکینسون و ریانی ، ۲۰۰۰؛ هاوکینس و همکاران ، ۱۹۸۴؛ ویس و لویس ، ۱۹۹۴). بر این اساس پس از شناسایی متغیرها و میزان تاثیرگذاری آن‌ها بر یکدیگر، به مدل‌سازی رابطه بین متغیرها با بیشترین دقت و کارایی پرداخته می‌شود (سالینی و همکاران ، ۲۰۱۶). اغلب روش‌های برازش پرکاربرد متداول مانند روش مجموع کمترین توان‌های دوم خطا و روش مجموع کمترین قدرمطلق انحرافات به شدت تحت تاثیر حتی یک مشاهده دورافتاده (مخصوصا نقاط اهرمی) هستند. اینگونه رویکردها قادر به شناسایی مشاهدات دورافتاده نیستند (بکر و قادر ، ۱۹۹۹؛ هاوکینس ، ۱۹۸۰؛ ویس و لویس ، ۱۹۹۴) و در حضور مشاهدات دورافتاده نتایج غیر قابل اعتماد و دور از واقعیت دارند (درویلیس و همکاران ، ۲۰۱۵). در این‌خصوص، روش‌های متنوعی در تحلیل مدل‌های رگرسیون استوار در حضور داده دورافتاده ارائه شده است (هوبرت و همکاران ، ۲۰۰۸؛ گوین و ولش ، ۲۰۱۰). این رویکردها عمدتا بر مبنای استفاده از توابع هدفی هستند که تاثیرات نامناسب مشاهدات دورافتاده را در برآورد مدل کم اثر می‌سازند (بیلور و همکاران ، ۲۰۰۶؛ روسو و لروی ، ۱۹۸۷). به عنوان نمونه تابع هدف کمترین قدرمطلق انحرافات نسبت به مشاهدات دورافتاده عمودی استوار است ولی در مقابل نقاط دورافتاده اهرمی عملکرد ضعیفی دارد. در این خصوص رگرسیون‌های مبتنی بر M-برآوردها مشاهدات دورافتاده اهرمی را کم اثر می‌کنند (هوبر و رونچتی ، ۲۰۰۹؛ ماروبینی و اورنتی ، ۲۰۱۴؛ روسو و وانزومن ، ۱۹۹۰). لذا بهتر است در مسئله بهینه‌سازی توابع هدف مناسبی در نظر گرفته شوند که متناسب با ماهیت و نوع مشاهدات دورافتاده عملکرد بهینه‌ای داشته باشند (ماروبینی و اورنتی ، ۲۰۱۴). به این منظور رویکردهای وزنی در برآورد مدل‌های رگرسیون کمترین قدرمطلق انحرافات معرفی شد. محمودی (۱۳۸۴) به رگرسیون حداقل قدرمطلق انحرافات وزنی استوار و لینگ (۲۰۰۵) به بررسی اینگونه برآوردگرها در مدل‌های اتو-رگرسیو پرداخت. گائو و فنگ (۲۰۱۸) رگرسیون کمترین قدرمطلق انحرافات وزنی تاوانیده را با هدف شناسایی داده‌های دورافتاده و معرفی برآوردگرهای استوار معرفی کرد. لازم به ذکر است که در رویکردهای معرفی شده در برآورد مدل‌های رگرسیونی کمترین قدرمطلق انحرافات وزنی، پارامترهای مدل بر اساس بهینه‌سازی تابع هدف $\sum_{i=1}^{n} w_i|e_i|$ برآورد می‌شوند. در این مقاله یک مدل رگرسیونی طبق شرایط زیر برآورد می‌شود:

۱- مسئله بهینه‌سازی با تابع هدفی بر مبنای الحاق وزنی قدرمطلق انحرافات مرتب شده $\sum_{i=1}^{n} w_i|e_{(i)}|$

در برازش مدل در نظر گرفته شود، که در آن $e_{(i)}$، i-امین بزرگترین مقدار خطا است.

۲- بهترین تابع هدف در برازش به داده‌ها در تناسب با تعداد و نوع مشاهدات دورافتاده به گونه‌ای تعیین شود که در برازش به مدل دارای بهینه‌ترین و بالاترین مقدار نقطه شکست [1] باشد. نقطه شکست یک برآوردگر کمترین تعداد مشاهداتی است که اگر مقادیر نامتعارف یا دورافتاده اختیار کنند مقدار برآوردگر را به شدت تغییر دهند (روسو و هوبرت ، ۲۰۱۳؛ هاوکینس و همکاران ، ۱۹۸۴). به عنوان مثال نقطه شکست برآوردگرهای مجموع کمترین توان‌های دوم خطا و مجموع کمترین قدرمطلق انحرافات یک است، یعنی فقط یک مشاهده دورافتاده کافی است که مقدار تابع هدف مسئله بهینه‌سازی را به شدت تحت تاثیر قرار دهد (گوین و ولش ، ۲۰۱۰). از طرفی نقطه شکست برآوردگر کمترین توان دوم میانه‌ی خطا $\frac{n}{۲}$ است یعنی $\frac{n}{۲}$ مشاهدات باید مقادیر دورافتاده اختیار کنند تا میانه تحت تاثیر قرار گیرد (روسو ، ۱۹۸۴).

۳- به تحلیل داده‌ها و مقادیر خطاهای برازش به منظور شناسایی و تعیین نقاط دورافتاده بپردازد.

برای برقراری شرایط ۱ تا ۳، پس از مرتب کردن انحرافات، الحاق و جمع‌بندی آن‌ها با اعمال بردارهای وزنی صورت می‌گیرد. برای شناسایی مشاهدات دورافتاده وزن‌های بهینه در نظر گرفته می‌شوند تا مناسبترین تابع برازش مدل رگرسیونی با یک نقطه شکست بالا در مقابل دیگر برآوردگرهای کاندید حاصل شود (هوبرت و همکاران ، ۲۰۰۸).

در بخش ۲ رگرسیون کمترین قدرمطلق انحرافات مرتب شده وزنی معرفی می‌شود. در بخش ۳ الگوریتم برآوردیابی مدل بیان می‌شود. در بخش‌های ۴، ۵ و ۶ عملکرد مدل پیشنهادی در مثال‌های تحلیلی، شبیه‌سازی شده، کاربردی با داده‌های واقعی در مهندسی آب مورد بررسی و مطالعه قرار می‌گیرد. در بخش انتهایی به بحث و نتیجه‌گیری پرداخته می‌شود.

۲ رگرسیون کمترین قدرمطلق انحرافات مرتب شده وزنی

یک مدل آماری به صورت مجموع تابعی از متغیرهای تبیینی و مقادیر خطا به صورت

$$\boldsymbol{y} = f_{\boldsymbol{\alpha}}(\boldsymbol{x}) + \boldsymbol{\varepsilon}, \qquad (۱)$$

در نظر گرفته شود، که در آن $\boldsymbol{y}$ متغیر وابسته، $\boldsymbol{x}$ متغیر تبیینی، $\boldsymbol{\varepsilon}$ مقادیر خطا با شرایط زیربنایی هستند و $f_{\boldsymbol{\alpha}}(\cdot)$ با پارامترهای $\boldsymbol{\alpha} = (\alpha_۰, \ldots, \alpha_p)$ رابطه بین متغیرها را مشخص می‌کند. انتخاب مدل مناسب

[1] Break-down point

شامل تعیین نوع تابع $f_{\alpha}(\cdot)$ و انتخاب روش برآوردیابی پارامترهای مدل است. پس از جمع‌آوری داده‌ها و انتخاب مدل رگرسیون خطی، به محاسبات آماری برای برآورد پارامترهای مدل و بررسی نکویی برازش مدل پرداخته می‌شود. فرض کنید برای مشاهدات $(y_1, x_1), \ldots, (y_n, x_n)$ رابطه خطی $y = \alpha_0 + \alpha_1 x + \varepsilon$ برقرار باشد. برای برآورد پارامترهای α_0 و α_1 یک مسئله بهینه‌سازی با تابع هدف $\mathcal{O} = [O_1, \ldots, O_n]$ در نظر گرفته می‌شود، که در آن $O_i = e_{(i)}$، $i = 1, \ldots, n$ به ترتیب هرکدام iامین قدرمطلق انحرافات مرتب شده را کمینه خواهند کرد، به طوری که

الف- $e_i = |y_i - (\alpha + \beta \times x_i)|$ قدرمطلق مقادیر انحرافات برای $i = 1, \ldots, n$ است،

ب- $e_{(1)} \leq \ldots \leq e_{(n)}$ مقادیر مرتب شده قدرمطلق انحرافات هستند.

بر خلاف مسایل بهینه‌سازی معمولی که در آن‌ها یک تابع هدف منفرد به منظور یافتن جوابها بهینه می‌شود، بیش از یک تابع هدف در بهینه‌سازی وارد می‌شود (کوردوس و همکاران ، ۲۰۱۹). در این مسئله بهینه‌سازی ابتدا کلیه توابع هدف توسط عملگرهایی (از جمله عملگرهای وزنی) به یک تابع تبدیل می‌شود (لیت و اسکریانس ، ۲۰۱۹؛ یاری و چاجی ، ۲۰۱۲a،b). در این مقاله با برقراری ارتباط بین وزن‌ها و نقاط دورافتاده به تعیین بهینه بردار وزن پرداخته می‌شود. در ادامه از بردارهای وزنی در الحاق و جمع‌بندی مقادیر خطا در یک تابع هدف به صورت

$$\min_{\mathbf{W}} \min_{\alpha_0, \alpha_1} \sum_{i=1}^{n} \omega_i e_{(i)}, \quad \omega_i \in [0, 1], \quad \sum_{i=1}^{n} \omega_i = 1, \tag{۲}$$

استفاده می‌شود، که در آن وزن‌ها با رویکردهای متنوعی (چاجی ، ۲۰۱۷؛ چاجی و همکاران ، ۲۰۱۸) قابل تعیین هستند. برای شناسایی مجموعه نقاط دورافتاده، مجموعه نقاط خوب و تحلیل حساسیت آستانه تمایز بین این دو مجموعه، بردارهای وزن به صورت زیر در نظر گرفته می‌شوند

$$\mathbf{W}_k = (\underbrace{0, \ldots, 0}_{\#k-1}, \underbrace{1}_{\#1}, \underbrace{0, \ldots, 0}_{\#n-k}), \quad k = [\frac{n}{2}] + 1, \ldots, n.$$

ملاحظه ۱. در رابطه (۲) خطاهای مرتب شده با اوزان مربوطه، تابع هدف مسئله بهینه‌سازی را تشکیل می‌دهند. منطقی است خطاهای بزرگ مربوط به مشاهداتی باشند که توسط مدل به خوبی برازش نشده‌اند. به این مشاهدات باید برچسب دورافتاده یا نامتعارف اختصاص داد. در نقطه مقابل مشاهداتی هستند که توسط مدل به خوبی برازش شده‌اند و دارای خطاهای برازشی کوچک هستند. به این مشاهدات برچسب

مشاهدات خوب اختصاص می‌دهیم. اکنون، باید مرز بین مشاهدات دورافتاده و خوب، تعیین شود. در اینجا مقدار k در بردار وزن $\mathbf{W}_k$ نقش تعیین کننده مقدار آستانه‌ای بین مجموعه مشاهدات خوب و مجموعه مشاهدات دورافتاده را دارا است. مقدار بهینه k براساس معیاری که فاصله بزرگترین مقدار خطای برازش مدل بر مجموعه داده‌های خوب را با کوچکترین مقدار خطای برازش مدل بر مجموعه داده‌های دورافتاده بیشینه می‌کند، تعیین می‌شود. بدیهی است که با توجه به مرتب بودن خطاهای برازشی، هر چه این فاصله بزرگتر باشد، مدل بهینه تمایز آشکارتری بین دو مجموعه داده‌های خوب و دورافتاده ایجاد می‌کند.

ملاحظه ۲. در یک مجموعه مشاهدات حداکثر نیمی از آن‌ها را می‌توان به عنوان داده دورافتاده در نظر گرفت. لذا با تحلیل حساسیت مقادیر فاصله خطاهای برازش بین مجموعه داده‌های خوب و دورافتاده بر اساس کلیه مقادیر $k = \left[\frac{n}{۲}\right] + ۱, \ldots, n$، بهترین مقدار k انتخاب می‌شود. جزییات نحوه محاسبه این فاصله و تعیین بهینه مقدار k در الگوریتمی که در ادامه بیان می‌شود، توضیح داده شده است.

با جایگذاری بردارهای $\mathbf{W}_k$ در (۲)، مسئله بهینه‌سازی به صورت

$$\min_{\left[\frac{n}{۲}\right]+۱\leq k\leq n} \min_{(\alpha_۰,\alpha_۱)} e_{(k)}, \qquad (۳)$$

بازنویسی می‌شود و برای حل آن ابتدا مسئله‌های بهینه‌سازی

$$(\widehat{\alpha}_{۰k}, \widehat{\alpha}_{۱k}) = \text{argmin}\ e_{(k)}, \quad k = \left[\frac{n}{۲}\right] + ۱, \ldots, n. \qquad (۴)$$

در نظر گرفته می‌شود. نقطه شکست در هر یک از مسئله‌های بهینه‌سازی (۴) برابر $n-k$ است، زیرا اگر $n-k$ مشاهده، دورافتاده باشند، مقادیر خطاهای مرتب شده متناظر با آن‌ها به ترتیب در مکان‌های $k+۱, \ldots, n$ به صورت $e_{(k+۱)}, \ldots, e_{(n)}$ قرار می‌گیرند. در نتیجه این مقادیر در مسئله بهینه‌سازی وارد نمی‌شوند و مقادیر پارامترهای برآورد شده را تحت تاثیر قرار نمی‌دهند. در این‌خصوص نیاز است با جستجو در $n - \left[\frac{n}{۲}\right]$ مسئله بهینه‌سازی (۴) مقدار بهینه k، یعنی k^*، به درستی تعیین شود. بر این اساس مناسبترین تابع نیکویی برازش برای برآوردیابی مدل به صورت $\min_{(\alpha_۰,\alpha_۱)} e_{(k^*)}$ نتیجه می‌شود. در نهایت با توجه به محاسبات مسئله‌های بهینه‌سازی (۴)، جواب مسئله بهینه‌سازی (۳)، به صورت $(\widehat{\alpha}_۰, \widehat{\alpha}_۱) = (\widehat{\alpha}_{۰k^*}, \widehat{\alpha}_{۱k^*})$ حاصل می‌شود. توجه کنید که در هر مسئله مقدار بهینه k^* متناسب با تعداد مشاهدات دورافتاده در مجموعه داده‌ها مشخص می‌شود. این مقدار مناسبترین تابع نیکویی برازش

برای برآوردیابی مدل را در هر مسئله نتیجه می‌دهد و از طرفی تعیین کننده بهترین نقطه شکست در مسئله بهینه‌سازی است که برابر $n - k^*$ است (هوبرت و همکاران ، ۲۰۰۸؛ روسو و هوبرت ، ۲۰۱۳).

ملاحظه ۳. بر اساس روش معرفی شده، هر انتخاب دلخواه و مناسب دیگری برای $\mathbf{W}_k$ها بطور مشابه قابل بررسی است و مدل جدیدی با خصوصیات متفاوتی را معرفی خواهد نمود.

۳ برآورد پارامترهای مدل

در یک مدل رگرسیون مبتنی بر کمترین قدرمطلق انحرافات، پارامترهای بهینه مدل بر اساس یک مسئله بهینه‌سازی خطی محاسبه می‌شوند. در این‌گونه مدل‌ها اثبات می‌شود که مدل برازش شده حداقل از $p+1$ نقطه از مشاهدات عبور می‌کند، که $p+1$ تعداد پارامترهای برآورد شده در مدل است. لذا در یک مدل رگرسیون مبتنی بر مسئله بهینه‌سازی (۳) ابتدا کلیه ترکیب‌های مختلف $p+1$تایی از n جفت مشاهدات $(y, \mathbf{x})$ انتخاب می‌شود و معادله صفحه گذرا از این نقاط به همراه میزان خطاهایی که دیگر نقاط با صفحه گذرا بوجود می‌آورند، مشخص می‌شود (اوگوندل و همکاران ، ۲۰۱۶). در انتها مدل بهینه در بین $\binom{n}{p+1}$ مدل کاندید و بر اساس مجموعه‌ای از معیارهای برازش موجود، انتخاب می‌شود.

الگوریتم ۱. الگوریتم برآوردیابی ضرایب مدل:

گام ۱- قرار دهید $k = [\frac{n}{2}] + 1$ و $\mathcal{S} = \{[(y_m, x_m), (y_\ell, x_\ell)] | m, \ell = 1, \ldots, n, \quad m < \ell\}$.

گام ۲- مدل‌های برازش کاندید را با انجام موارد زیر تعیین کنید:

۱- معادله خط گذرا از دو جفت نقطه دلخواه $[(y_m, x_m), (y_\ell, x_\ell)]$ در گردایه $\mathcal{S}$ با شیب a و عرض از مبدا b را بدست آورید.

۲- قدرمطلق انحرافات کل مشاهدات را به صورت $e_i = |y_i - a - bx_i|$ بدست آورید.

۳- مقادیر $e_{(1)} \leq \ldots \leq e_{(n)}$ را محاسبه کنید و مقدار $e_{(k)}$ را تعیین نمایید.

۴- موارد ۱ تا ۳ را برای تمام جفت نقاط موجود در $\mathcal{S}$ انجام دهید.

۵- معادله خط گذرا از جفت نقطه‌ای که کمترین مقدار $e_{(k)}$ را در بین کلیه اعضای $\mathcal{S}$ دارد را به عنوان بهترین خط برازش مرحله kام انتخاب نموده، قرار دهید $(\widehat{a}_k, \widehat{b}_k) = \text{argmin}_{a,b} e_{(k)}$.

توجه شود که در این مرحله $(\widehat{a}_k, \widehat{b}_k)$ می‌تواند به عنوان یک برآورد بهینه برای مدل‌سازی مشاهدات $(y_1, x_1), \ldots, (y_n, x_n)$، استفاده شود.

۶- قدرمطلق انحرافات نسبت به مدل برآورد شده مرحله kام و خطای نسبی هر مشاهده را به صورت

$$E_i = |y_i - (\widehat{a}_k + \widehat{b}_k \times x_i)|, \quad \delta_i = \frac{E_{(i)}}{\sum_{i=1}^{n} e_i}, \quad i = 1, \ldots, n,$$

محاسبه کنید، که در آن $E_{(1)} \leqslant \ldots \leqslant E_{(n)}$. توجه کنید که مقادیر بزرگ δ_i نشانگر برازش نامناسب مدل به مشاهده iام است. مقدار $\Delta_k = \delta_{k+1} - \delta_k$ آستانه بین نیکویی برازش مدل به داده‌های خوب و بد را اندازه‌گیری می‌کند. هر چه این مقدار برای یک مدل بزرگتر باشد، نشان می‌دهد که آن مدل توانایی بهتری در شناسایی و جداسازی مشاهدات دورافتاده دارد.

گام ۳- قرار دهید $k = k + 1$ اگر $k \leq n$ آنگاه گام ۲ را تکرار کنید.

گام ۴- (انتخاب مدل بهینه): در بین کلیه مدل‌های انتخاب شده در مرحله ۶ از گام ۲ با مقادیر Δ_k، مقدار بهینه اندیس k را به‌صورت $k^* = \text{argmax}_{k=n-[\frac{n}{2}],\ldots,n}\Delta_k$ انتخاب کنید. عدد k^* مربوط به مدلی است که در بین مدل‌های کاندید در مرحله ۶ از گام ۲ بیشترین مقدار اختلاف در مقادیر نیکویی برازش به مشاهدات خوب و دورافتاده را اعمال می‌کند. برآوردهای مدل مربوط به این مدل، یعنی $(\widehat{\alpha}_0, \widehat{\alpha}_1) = (\widehat{a}_{k^*}, \widehat{b}_{k^*})$، به همراه تابع نیکویی برازش $e_{(k^*)}$ مربوطه را به عنوان مدل بهینه و بهترین تابع نیکویی برازش انتخاب کنید و قرار دهید $\widehat{y}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 x_i$. مقدار $n - k^*$ بهترین نقطه شکست برآوردگر مدل بهینه را مشخص می‌کند. هرچه مقدار k^* بزرگتر باشد، تعداد مشاهداتی که توسط مدل به خوبی برازش می‌شوند، بیشتر هستند و به تبع آن مشاهدات کمتری توسط مدل به خوبی برازش نمی‌شوند و به عنوان مشاهدات دورافتاده شناسایی می‌شوند.

گام ۵- (تعیین نقاط دورافتاده): اگر مشاهدات شامل نقطه یا نقاط دورافتاده باشند، مقدار δ_{k^*} تفاوت و پرش قابل ملاحظه‌ای با مقدار بعد از خود یعنی δ_{k^*+1} دارد. k^* را به عنوان تعداد مشاهدات خوب مشخص کنید. اکنون قدرمطلق خطاهای برازش مدل بهینه یعنی $e_i = |y_i - (\widehat{\alpha}_0 + \widehat{\alpha}_1 \times x_i)|$ را برای هر $i = 1, \ldots, n$ محاسبه نمایید. با جستجو در مجموعه مشاهدات، آن دسته از مشاهداتی که مقادیر $\delta_{k^*+1}, \ldots, \delta_n$ را تولید می‌کنند به عنوان مشاهدات دورافتاده گزارش کنید.

۴ مطالعه تحلیلی

برای بررسی و مقایسه رویکرد معرفی شده در مقایسه با برآوردگرهای متداول کمترین قدرمطلق انحرافات (LA) و کمترین توان‌های دوم خطا (LS) مشاهدات جدول ۱ را به صورت

$$x_i \sim \text{Uniform}(۰, ۳), \quad y_i = ۲ + ۴x_i + N(۰, ۱), \quad i = ۱, \ldots, ۸,$$
$$x_i \sim \text{Uniform}(۳, ۴), \quad y_i = x_i - \text{Uniform}(۱, ۲), \quad i = ۹, ۱۰.$$

شبیه‌سازی شده‌اند. در جدول‌های ۱ و ۲ بر مبنای گام ۳ الگوریتم ۱، نتایج مدل‌های بهینه برای توابع هدف

جدول ۱. مشاهدات مطالعه تحلیلی

i	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
x_i	۲٫۵	۲٫۸	۲٫۱	۰٫۸	۰٫۳	۰٫۸	۱٫۱	۰٫۶	۳٫۸	۳٫۶
y_i	۱۱	۱۳	۹	۷	۴	۴	۵	۵	۲	۲

$e_{(k)}$، $k = ۶, \ldots, ۱۰$ و دو مدل کمترین قدرمطلق انحرافات LA و کمترین توان‌های دوم خطا LS آورده شده است. توجه کنید که در این مرحله هر یک از مدل‌های موجود در این جداول، خود به تنهایی می‌تواند به عنوان یک مدل مناسب برای مدل‌سازی این مشاهدات بکار برده شود. اکنون مطابق گام ۴ الگوریتم، به انتخاب بهترین مدل بهینه در بین مدل‌های کاندید جدول ۱ پرداخته می‌شود. به این منظور نتایج جدول ۲ در جهت انتخاب و معرفی مدل بهینه برای این مشاهدات به صورت زیر تفسیر می‌شوند.

۱- در مدل با $k = ۶$ چهار مشاهده پتانسیل دورافتاده بودن را دارا خواهند بود. زیرا در این مدل ۸۲٫۶ درصد از خطای کل ($\delta_۷ + \delta_۸ + \delta_۹ + \delta_{۱۰} = ۰٫۸۲۶$) مربوط به چهار مشاهده است (مشاهدات دورافتاده) و ۱۷٫۴ درصد باقیمانده مربوط به شش مشاهده دیگر است (مشاهدات خوب). در این مدل مشاهداتی که مقادیر $\delta_۷$، $\delta_۸$، $\delta_۹$، $\delta_{۱۰}$ را تولید می‌کند، مشاهده دورافتاده هستند.

۲- در مدل با $k = ۷$ سه مشاهده پتانسیل دورافتاده بودن را دارند. در این مدل ۸۲٫۱ درصد از خطای کل مربوط به سه مشاهده است (مشاهدات دورافتاده) و ۱۷٫۹ درصد باقیمانده مربوط به هفت مشاهده دیگر است (مشاهدات خوب). بنابراین باید سه مشاهده مربوط به $\delta_۸$، $\delta_۹$، $\delta_{۱۰}$ را به عنوان مشاهدات دورافتاده تعیین نمود. پس مشاهده‌ای که مقدار خطای نسبی $\delta_۸ = ۰/۰۷$ را تولید می‌کند به عنوان مشاهده دورافتاده معرفی می‌شود که مقدار خطای نسبی آن در مقایسه با $\delta_۹ = ۰٫۳۷$ و $\delta_{۱۰} = ۰٫۳۸$ تفاوت زیادی

دارد. در صورت انتخاب این مدل یکی از مشاهداتی که برازش خوبی از مدل دریافت می‌کند باید به عنوان یک مشاهده دورافتاده معرفی شود. البته مقدار $\Delta_۷ = ۰٫۰۲$ تاکید می‌کند که این مدل تمایز معنی‌داری در مقادیر نیکویی برازش بین مشاهدات دورافتاده و خوب ایجاد نمی‌کند.

۳- مدل $k = ۸$ که از کمینه کردن تابع برازش $\min e_{(۸)}$ به دست می‌آید، در مقایسه با مدل‌های دیگر مدل برتر است. در این مدل ۸۰ درصد از خطای کل ($\delta_۹ + \delta_{۱۰} = ۰٫۸۰$) مربوط به دو مشاهده است و ۲۰ درصد دیگر مربوط به هشت مشاهده باقیمانده است. از طرفی تفاوت معنی‌داری بین مقدار $\delta_۸ = ۰٫۰۵$ و $\delta_۹ = ۰٫۳۹$ است. لذا مشاهده مربوط به $\delta_۸$ و همچنین مشاهدات مربوط به $\delta_۱, \ldots, \delta_۷$ از برازش خوبی برخوردار هستند و نمی‌تواند به عنوان مشاهدات دورافتاده معرفی شوند. به عبارتی تفاوت معنی‌داری بین خطاهای برازش مدل به مشاهدات خوب و دورافتاده وجود دارد ($\Delta_۸ = ۰٫۳۴$).

۴- مدل با $k = ۹$ نمی‌تواند به عنوان مدل بهینه انتخاب شود، زیرا در این مدل می‌توان یک مشاهده را قطعا به عنوان مشاهده دورافتاده معرفی نمود که حداکثر خطای نسبی تولیدی آن ۲۹ درصد است. بدیهی است که ۷۱ درصد از خطای نسبی تولید شده توسط این مدل در ۹ مشاهده پخش شده است.

۵- در مدل‌های دیگر جدول ۱ تفاوت معنی‌داری بین بزرگترین مقادیر خطاهای نسبی وجود ندارد. به عنوان نمونه مدل با $k = ۱۰$ در مقایسه با دیگر مدل‌ها، اختلاف‌های معنی‌داری بین بزرگترین مقادیر خطاهای نسبی بوجود نمی‌آورد. همچنین با توجه به ماهیت بهینه‌سازی مدل‌های LS و LA مقدار کل خطا بطور متوسط در بین کلیه مشاهدات انتشار می‌یابد. ماهیت انتشار خطا بطور متوسط باعث می‌شود که بخشی از مشاهدات به خوبی برازش نشوند و دارای خطای برازش نسبتا بزرگی باشند. این نتایج در شکل ۶ که نمودارهای میله‌ای مقادیر خطاهای نسبی مرتب شده مدل‌های جدول ۲ را نشان می‌دهد، قابل استنتاج است. لذا در این دو مدل نمی‌توان به تعیین مشاهدات دورافتاده پرداخت.

اکنون برای مدل بهینه $\widehat{y}_i = ۳٫۰۵ + ۳٫۱۸x_i$، مطابق گام ۵ الگوریتم ۱، مشاهدات دورافتاده تعیین می‌شوند. بردار خطاهای برازش این مدل در جدول ۴ بیانگر آن است که مشاهدات $(y_۹, x_۹) = (۲, ۳٫۸)$ و $(y_{۱۰}, x_{۱۰}) = (۲, ۳٫۶)$ به ترتیب بزرگترین مقادیر خطا ۱۳٫۱ و ۱۲٫۵ را تولید می‌کنند. این دو مقدار تفاوت زیادی با دیگر مقادیر خطا دارند و به عنوان مشاهدات دورافتاده شناسایی می‌شوند.

۵ مطالعه شبیه‌سازی و تحلیل حساسیت نتایج

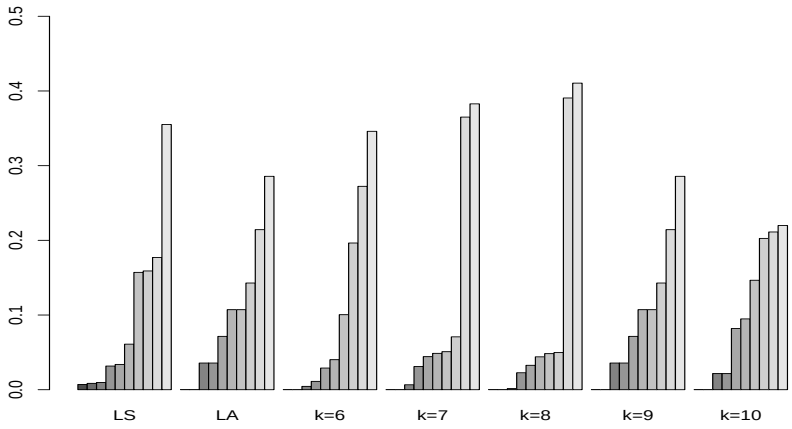
در این بخش عملکرد برآوردگرهای استوار پیشنهاد شده ($\widehat{\boldsymbol{\beta}}_{\text{rob.}}$) در مقابل برآوردگرهای کمترین توان دوم خطا ($\widehat{\boldsymbol{\beta}}_{\text{LS}}$) و برآوردگرهای کمترین قدرمطلق انحرافات ($\widehat{\boldsymbol{\beta}}_{\text{LA}}$) مقایسه می‌شوند. برای این منظور، از

جدول ۲. مدل‌های کاندید برآورد شده در گام ۳ الگوریتم به همراه مقادیر خطاهای مرتب شده

$e_{(۱۰)}$	$e_{(۹)}$	$e_{(۸)}$	$e_{(۷)}$	$e_{(۶)}$	$e_{(۵)}$	$e_{(۴)}$	$e_{(۳)}$	$e_{(۲)}$	$e_{(۱)}$	$\sum_{i=۱}^{n=۱۰} e_i$	(a_k, b_k)	$\min e_{(k)}$	k
۱۰٫۳	۸٫۱	۵٫۹	۳٫۰	۱٫۲	۰٫۸۷	۰٫۳۳	۰٫۱۳۳	۰	۰	۳۰	$(۴٫۵۳, -۰٫۶۷)$	$\min e_{(۶)}$	۶
۱۱٫۵	۱۱٫۰	۲٫۱	۱٫۵	۱٫۵	۱٫۳۳	۰٫۹۳	۰٫۲۰۰	۰	۰	۳۰	$(۳٫۴۰, ۲٫۶۷)$	$\min e_{(۷)}$	۷
۱۳٫۱	۱۲٫۵	۱٫۶	۱٫۵	۱٫۴	۱٫۰۵	۰٫۷۳	۰٫۰۴۵	۰	۰	۳۲	$(۳٫۰۵, ۳٫۱۸)$	$\min e_{(۸)}$	۸
۸٫۰	۶٫۰	۴٫۰	۳٫۰	۳٫۰	۲٫۰۰	۱٫۰۰	۱٫۰۰۰	۰	۰	۲۸	$(۵٫۰۰, ۰٫۰۰)$	$\min e_{(۹)}$	۹
۶٫۴	۶٫۱	۵٫۹	۴٫۲	۲٫۸	۲٫۳۸	۰٫۶۲	۰٫۶۲۵	۰	۰	۲۹	$(۳٫۶۲, ۱٫۲۵)$	$\min e_{(۱۰)}$	۱۰
۸٫۰	۶٫۰	۴٫۰	۳٫۰	۳٫۰	۲٫۰۰	۱٫۰۰	۱٫۰۰۰	۰	۰	۲۸	$(۵٫۰۰, ۰٫۰۰)$	$\min \sum \lvert e_i \rvert$	LA
۴۴٫۵	۲۲٫۲	۱۹٫۹	۱۹٫۶	۷٫۶	۴٫۲۴	۳٫۹۶	۱٫۲۱۰	۱٫۰۷	۰٫۸۸	۱۲۵٫۳۲	$(۵٫۵۹, ۰٫۱۳)$	$\min \sum e_i^۲$	LS

جدول ۳. خطاهای نسبی مرتب شده مدل‌های کاندید بر اساس گام ۴ الگوریتم برای انتخاب مدل بهینه

Δ_k	$\sum_{i=k+۱}^{۱۰} \delta_i$	$\sum_{i=۱}^{k} \delta_i$	$\delta_{۱۰}$	$\delta_۹$	$\delta_۸$	$\delta_۷$	$\delta_۶$	$\delta_۵$	$\delta_۴$	$\delta_۳$	$\delta_۲$	$\delta_۱$	k
۰٫۰۶	۰٫۸۲۶	۰٫۱۷۴	۰٫۳۵	۰٫۲۷	۰٫۱۹۶	۰٫۱۰۰	۰٫۰۴۰	۰٫۰۲۹	۰٫۰۱۱	۰٫۰۰۴	۰	۰	۶
۰٫۰۲	۰٫۸۲۱	۰٫۱۷۹	۰٫۳۸	۰٫۳۷	۰٫۰۷۱	۰٫۰۵۱	۰٫۰۴۹	۰٫۰۴۴	۰٫۰۳۱	۰٫۰۰۶	۰	۰	۷
۰٫۳۴	۰٫۸۰	۰٫۲۰	۰٫۴۱	۰٫۳۹	۰٫۰۵۰	۰٫۰۴۸	۰٫۰۴۴	۰٫۰۳۳	۰٫۰۲۲	۰٫۰۰۱	۰	۰	۸
۰٫۰۸	۰٫۲۹	۰٫۷۱	۰٫۲۹	۰٫۲۱	۰٫۱۴۳	۰٫۱۰۷	۰٫۱۰۷	۰٫۰۷۱	۰٫۰۳۵	۰٫۰۳۵	۰	۰	۹
–	۰٫۰۰	۱٫۰۰	۰٫۲۲	۰٫۲۱	۰٫۲۰۳	۰٫۱۴۷	۰٫۰۹۵	۰٫۰۸۲	۰٫۰۲۱	۰٫۰۲۱	۰	۰	۱۰
–	–	–	۰٫۲۹	۰٫۲۱	۰٫۱۴۳	۰٫۱۰۷	۰٫۱۰۷	۰٫۰۷۱	۰٫۰۳۵	۰٫۰۳۵	۰	۰	LA
–	–	–	۰٫۳۵	۰٫۱۸	۰٫۱۵۹	۰٫۱۵۷	۰٫۰۶۱	۰٫۰۳۴	۰٫۰۳۱	۰٫۰۰۹	۰٫۰۰۸	۰٫۰۰۷	LS



شکل ۱. نمودارهای میله‌ای مقادیر خطاهای نسبی مدل‌های جدول ۲

جدول ۴. خطاهای برازش مدل بهینه

i	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
e_i	۰	۱٫۰۵	۰٫۷۳	۱٫۴	۰	۱٫۶	۱٫۵	۰٫۰۴۵	۱۳٫۱	۱۲٫۵

مقیاس‌های وسیع‌تری از مدل‌های شبیه‌سازی شده استفاده می‌شود، که هر یک از نقطه نظر حجم نمونه، انواع و تعداد مشاهدات دورافتاده و تعداد پارامترها یا تعداد متغیرهای ورودی متفاوت هستند. حال به تحلیل حساسیت نتایج در ارتباط با واریانس خطا $\sigma_e^۲$، حجم نمونه n (شامل $n_۱$ مشاهده خوب و $n_۲ = n - n_۱$ مشاهده دورافتاده)، متوسط مقادیر برآورد شده برای پارامترها، انحراف معیار و کارایی برآوردگرها پرداخته می‌شود. برای ارایه نتایج کاملی پیرامون مقایسه بین رویکرد پیشنهادی با رویکردهای کمترین توان دوم خطا و کمترین قدرمطلق انحرافات، هر یک از این شبیه‌سازی‌ها به تعداد $M = ۱۰۰۰۰$ بار تکرار شده‌اند. نتایج مربوط به مقادیر و شاخص‌های زیر برای هر مدل برازش شده محاسبه می‌شود:

$$\widehat{\sigma_e^۲}(m) = \frac{۱}{n-۱}\sum_{i=۱}^{n}(e_i(m)-\bar{e}(m))^۲, \quad m = ۱,\ldots,M,$$

$$\overline{\widehat{\sigma_e^۲}} = \frac{۱}{M}\sum_{m=۱}^{M}\widehat{\sigma_e^۲}(m), \qquad s^۲_{\widehat{\sigma_e^۲}} = \frac{۱}{M}\sum_{m=۱}^{M}(\widehat{\sigma_e^۲}(m)-\overline{\widehat{\sigma_e^۲}})^۲,$$

$$
\begin{aligned}
\overline{\widehat{\beta}_j} &= \frac{1}{M}\sum_{m=1}^{M}\widehat{\beta}_j(m), \qquad s^2_{\widehat{\beta}_j} = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\beta}_j(m)-\overline{\widehat{\beta}_j})^2,\\
RMSE(\widehat{\beta}_j) &= \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\beta}_j(m)-\beta_j)^2}, \quad j=0,1,\ldots,p,\\
\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}},\widehat{\boldsymbol{\beta}}_{\text{LA}}) &= \frac{\sum_{m=1}^{M}||\widehat{\boldsymbol{\beta}}_{\text{LA}}(m)-\boldsymbol{\beta}||_2^2}{\sum_{m=1}^{M}||\widehat{\boldsymbol{\beta}}_{\text{rob.}}(m)-\boldsymbol{\beta}||_2^2},\\
\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}},\widehat{\boldsymbol{\beta}}_{\text{LS}}) &= \frac{\sum_{m=1}^{M}||\widehat{\boldsymbol{\beta}}_{\text{LS}}(m)-\boldsymbol{\beta}||_2^2}{\sum_{m=1}^{M}||\widehat{\boldsymbol{\beta}}_{\text{rob.}}(m)-\boldsymbol{\beta}||_2^2},\\
\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{LS}},\widehat{\boldsymbol{\beta}}_{\text{LA}}) &= \frac{\sum_{m=1}^{M}||\widehat{\boldsymbol{\beta}}_{\text{LA}}(m)-\boldsymbol{\beta}||_2^2}{\sum_{m=1}^{M}||\widehat{\boldsymbol{\beta}}_{\text{LS}}(m)-\boldsymbol{\beta}||_2^2},
\end{aligned}
$$

که در آن $\boldsymbol{e}(m)=(e_1(m),\ldots,e_n(m))$، $||\cdot||_2^2$ نرم اقلیدسی، $\boldsymbol{\beta}=(\beta_0,\ldots,\beta_p)$ پارمترهای مدل شبیه‌سازی شده و $\text{eff}(\cdot,\cdot)$ کارایی متقابل دو بردار از برآوردگرها است.

۵.۱ شبیه‌سازی ۱

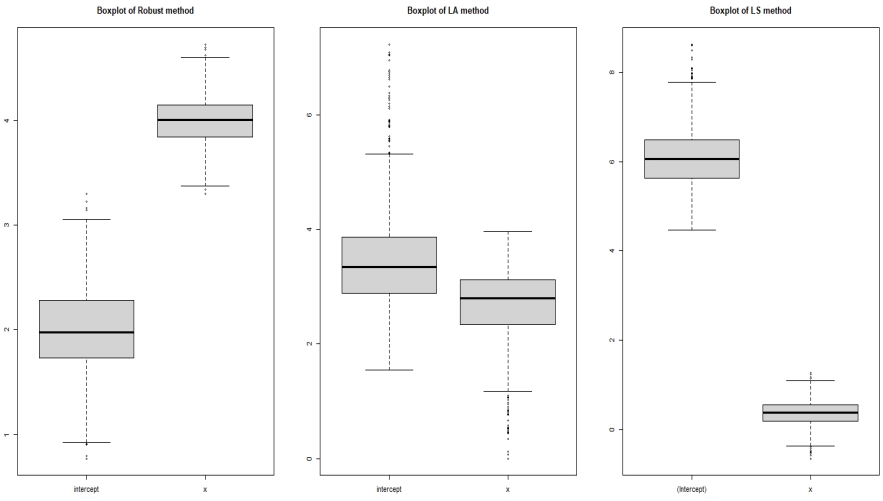
نتایج مقایسه رویکردهای موجود برای مدل

$$
\begin{aligned}
&x_i \sim \text{Uniform}(0,3), \quad y_i = 2+4x_i+N(0,1), \quad i=1,\ldots,n_1,\\
&x_i \sim \text{Uniform}(3,4), \quad y_i = x_i-\text{Uniform}(1,2), \quad i=n_1+1,\ldots,n.
\end{aligned}
$$

در جدول ۵ ارائه شده است. همچنین نمودارهای جعبه‌ای مقادیر پارامترهای برآورد شده این مدل‌ها در شکل ۲ برای حالتی که $n=50$ و تعداد مشاهدات خوب $n_1=40$ و تعداد مشاهدات دورافتاده $n_2=10$ است، نشان داده شده است.

۱- به نظر می‌رسد برآوردگرهایی که روش پیشنهادی در این مقاله معرفی می‌کند، ویژگی ناارریبی را داشته باشند. اگرچه که این ویژگی باید در یک مطالعه نظری مورد بررسی قرار گیرد، اما بر اساس نتایج ارایه شده در سطر دوم جدول ۵ و جدول ۶ این موضوع دور از انتظار نیست. در این سطر میانگین مقادیر پارامترهای برآورد شده در ۱۰۰۰۰ تکرار برابر مقادیر اصلی پارامترها است.

۲- مقادیر $RMSE$ برای برآوردگرهای استوار در مقایسه با برآوردگرهای کمترین توان دوم خطا و کمترین



(ج) (ب) (الف)

شکل ۲. نمودار برآورد پارامترهای مدل الف- کمترین توان دوم خطا، ب- کمترین قدرمطلق انحرافات و ج- پیشنهاد شده

جدول ۵. نتایج مدل‌های برازش شده در ۱۰۰۰۰ تکرار شبیه‌سازی ($n = ۵۰$ و $n_۱ = ۴۰$)

پارامترها و شاخص‌ها	مدل‌ها $rob.$	LA	LS
$(\widehat{\sigma_e^۲}, s_{\widehat{\sigma_e^۲}})$	(۳۲٫۹, ۲٫۶)	(۲۲٫۹, ۴٫۳)	(۱۶٫۰, ۱٫۹)
$(\widehat{\beta_۰}, \widehat{\beta_۱})$	(۱٫۹۹, ۳٫۹۹)	(۳٫۴۸, ۲٫۵۹)	(۶٫۱۱, ۰٫۳۵)
$(s_{\widehat{\beta_۰}}, s_{\widehat{\beta_۱}})$	(۰٫۴۰, ۰٫۲۳)	(۰٫۸۸, ۰٫۷۷)	(۰٫۷۱, ۰٫۲۸)
$RMSE(\widehat{\boldsymbol{\beta}})$	(۰٫۴۰, ۰٫۲۳)	(۱٫۷, ۱٫۶)	(۴٫۲, ۳٫۶)
کارایی متقابل	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}\widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۱۳۹٫۳$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LA}}) = ۲۴٫۲$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{LA}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۵٫۴$

قدرمطلق انحرافات کمتر است.

۳- مقادیر انحراف معیار برآوردگرهای استوار کمتر از مقادیر انحراف معیار برآوردگرهای حاصل شده از دو رویکرد دیگر است. این مقادیر نشان می‌دهنده که برآوردگرهای استوار در مقایسه با دیگر برآوردگرها، کارا و از دقت بهتری در برآورد مقادیر پارامترها برخوردار است.

۴- بردار برآوردگرهای استوار در مقایسه با بردار برآوردگرهای کمترین توان دوم خطا و بردار کمترین قدرمطلق انحرافات کارا است.

جدول ۶. نتایج مقایسه مدل‌های برازش شده در ۱۰۰۰۰ تکرار شبیه‌سازی ($n = ۲۰۰$ و $n_۱ = ۱۵۰$)

پارامترها و شاخص‌ها	مدل‌ها $rob.$	LA	LS
$(\widehat{\sigma_e^۲}, s_{\widehat{\sigma_e^۲}})$	(۳۷٫۸, ۱٫۴۴)	(۱۷٫۴, ۲٫۴۸)	(۱۶٫۵, ۱٫۰۳)
$(\widehat{\beta}_۰, \widehat{\beta}_۱)$	(۱٫۹۹, ۳٫۹۹)	(۵٫۵۴, ۰٫۴۳)	(۶٫۵۱, −۰٫۰۰۰۷)
$(s_{\widehat{\beta}_۰}, s_{\widehat{\beta}_۱})$	(۰٫۲۰, ۰٫۱۱)	(۰٫۷۶, ۰٫۷۴)	(۰٫۳۷, ۰٫۱۳)
$RMSE(\widehat{\boldsymbol{\beta}})$	(۰٫۲۰, ۰٫۱۱)	(۳٫۶۲, ۳٫۶۴)	(۴٫۵۲, ۴٫۰۰)
کارایی متقابل	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۶۶۳٫۸$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LA}}) = ۴۸۰٫۳$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{LA}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۱٫۳۸$

جدول ۷. نتایج مدل‌های ۱ در مطالعه شبیه‌سازی ۲ در ۱۰۰۰۰ بار تکرار ($n = ۲۰۰$)

$n_۱$	پارامترها و شاخص‌ها	مدل‌ها $rob.$	LA	LS
	$(\widehat{\sigma_e^۲}, s_{\widehat{\sigma_e^۲}})$	(۵۱۶۸٫۶, ۴۱۸٫۳)	(۶۲۹٫۰, ۲۲۰٫۱)	(۴۳۰٫۲, ۸۰٫۱)
۱۹۰	$\widehat{\boldsymbol{\beta}}$	(۲٫۰۰, ۴٫۰۲, ۶٫۰۰, ۰٫۹۹)	(۹۷٫۰۳, ۴٫۳۱, ۵٫۴۷, ۰٫۲۳)	(۱۱۲٫۴۰, ۴٫۰۳, ۵٫۱۲, ۰٫۱۰)
	$\mathbf{s}_{\widehat{\boldsymbol{\beta}}}$	(۰٫۸۱, ۰٫۲۲, ۰٫۰۰۹, ۰٫۰۰۶)	(۹۵٫۵۱, ۳٫۰۵, ۰٫۵۴۵, ۰٫۷۷۰)	(۸٫۰۲, ۳٫۵۲, ۰٫۱۸۳, ۰٫۰۶۲)
	$RMSE(\widehat{\boldsymbol{\beta}})$	(۰٫۸۱, ۰٫۲۲, ۰٫۰۰۹, ۰٫۰۰۶)	(۱۰۳٫۸, ۳٫۸۳, ۱٫۱۰, ۰٫۸۴)	(۱۱۰٫۶۹, ۳٫۵۲, ۰٫۸۹۵, ۰٫۸۹۴)
	$Effi.$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۱۷۰۲۱$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LA}}) = ۱۲۶۷۳$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{LA}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۱٫۳۴$
	$(\widehat{\sigma_e^۲}, s_{\widehat{\sigma_e^۲}})$	(۱۷۴۶۲٫۸, ۷۴۴)	(۹۹۹٫۸, ۱۴۳)	(۸۴۶٫۶, ۹۱)
۱۶۰	$\widehat{\boldsymbol{\beta}}$	(۲٫۰۱, ۴٫۰۰, ۵٫۹۹, ۰٫۹۹)	(۱۰۵٫۶۴, ۴٫۳۶, ۴٫۹۰, ۰٫۱۶)	(۸۱٫۸۸, ۳٫۴۵, ۳٫۹۵, ۰٫۳۴)
	$\mathbf{s}_{\widehat{\boldsymbol{\beta}}}$	(۰٫۸۶, ۰٫۲۴, ۰٫۰۱, ۰٫۰۰۷)	(۵۹, ۳٫۸, ۰٫۱۶, ۰٫۴۴)	(۷٫۲, ۵٫۱, ۰٫۲۲, ۰٫۰۵)
	$RMSE(\widehat{\boldsymbol{\beta}})$	(۰٫۸۶, ۰٫۲۴, ۰٫۰۱, ۰٫۰۰۷)	(۱۰۳٫۸, ۳٫۸۳, ۱٫۱۰, ۰٫۸۴)	(۸۰٫۲۱, ۵٫۱۸, ۲٫۰۵, ۰٫۶۶)
	$Effi.$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۷۹۶۳$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LA}}) = ۱۳۲۹۵$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{LA}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۰٫۵۹$
	$(\widehat{\sigma_e^۲}, s_{\widehat{\sigma_e^۲}})$	۲۴۹۴۳٫۶, ۸۲۱٫۵)	۱۰۷۷٫۷, ۱۳۶٫۷)	(۹۷۴٫۴, ۸۲٫۴)
۱۳۰	$\widehat{\boldsymbol{\beta}}$	(۲٫۰۲, ۴٫۰۰, ۵٫۹۹, ۰٫۹۹)	(۶۳٫۴۲, ۴٫۴۲, ۳٫۳۶, ۰٫۴۸)	(۵۶٫۶۱, ۳٫۰۰, ۳٫۰۹, ۰٫۵۳)
	$\mathbf{s}_{\widehat{\boldsymbol{\beta}}}$	(۰٫۹۵, ۰٫۲۷, ۰٫۰۱, ۰٫۰۰۷)	(۲۵٫۵۸, ۵٫۰۴, ۰٫۸۷, ۰٫۱۹۵)	(۶٫۷۳, ۵٫۶۹, ۰٫۱۹, ۰٫۰۴۵)
	$RMSE(\widehat{\boldsymbol{\beta}})$	(۰٫۹۵, ۰٫۲۷, ۰٫۰۱, ۰٫۰۰۷)	(۶۶٫۵۳, ۵٫۰۶, ۲٫۷۷, ۰٫۵۵۴)	(۵۵٫۰۳, ۵٫۷۷, ۲٫۹۱, ۰٫۴۶۵)
	$Effi.$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۳۰۹۶٫۰۸$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LA}}) = ۴۴۹۸٫۰۶$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{LA}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۰٫۶۸$
	$(\widehat{\sigma_e^۲}, s_{\widehat{\sigma_e^۲}})$	(۲۷۲۸۹, ۸۷۱٫۷)	(۱۳۳۶, ۱۴۵٫۸)	(۹۵۳, ۸۰٫۱)
۱۱۰	$\widehat{\boldsymbol{\beta}}$	(۱٫۹۵, ۳٫۹۹, ۶٫۰۰, ۱٫۰۰)	(۲٫۲۷, ۱٫۴۳, ۱٫۱۹, ۰٫۹۶)	(۴۳٫۴۴, ۲٫۶۵, ۲٫۶۵, ۰٫۶۳)
	$\mathbf{s}_{\widehat{\boldsymbol{\beta}}}$	(۱٫۰۴, ۰٫۲۸, ۰٫۰۱, ۰٫۰۰۸)	(۵۵۲, ۱٫۴۷, ۰٫۲۰, ۰٫۰۴۰)	(۶٫۷۵, ۵٫۵۲, ۰٫۱۸, ۰٫۰۴۳)
	$RMSE(\widehat{\boldsymbol{\beta}})$	(۱٫۰۴, ۰٫۲۸, ۰٫۰۱, ۰٫۰۰۸)	(۵۵۳, ۲٫۹۶, ۴٫۸۱, ۰٫۰۵۲)	(۴۱٫۹۸, ۵٫۶۷, ۳٫۳۵, ۰٫۳۶۵)
	$Effi.$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۱۵۴۱٫۲$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{rob.}}, \widehat{\boldsymbol{\beta}}_{\text{LA}}) = ۵۳٫۳$	$\text{eff}(\widehat{\boldsymbol{\beta}}_{\text{LA}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}) = ۲۸٫۹$

۵- واریانس خطاهای به دست آمده در مدل پیشنهادی، بیشتر از واریانس خطای دو مدل دیگر است. این

موضوع در ابتدا به ماهیت برازش مدل‌های کمترین توان دوم خطا و کمترین قدرمطلق انحرافات که سعی بر کمینه‌سازی متوسط خطای هر مشاهده را دارند، مربوط است. توجه شود که در رویکرد پیشنهادی، هدف این است که تا حد امکان به مشاهدات خوب مدلی برازش شود که خطاهای اندکی داشته باشد. این موضوع باعث می‌شود که در مجموعه مقادیر خطاهای برازش، برخی از خطاها کوچک و با فاصله از خطاهایی باشند که از مشاهدات دورافتاده حاصل می‌شوند. هرچه این فاصله بیشتر باشد، تمایز آشکارتری در برازش مدل به مجموعه مشاهدات خوب و مجموعه مشاهدات دورافتاده حاصل می‌شود (به عنوان نمونه به نمودارهای میله‌ای در شکل ۶ برای $k = ۸$ و LA توجه کنید). لذا بدیهی است که پراکندگی چنین مقادیری نسبت به حالتی که خطاها بطور متوسط و میانگین توزیع شده‌اند، بیشتر باشند. از جهتی دیگر، باید در یک مطالعه نظری بررسی کرد که واریانس خطای مدل چه مقداری دارد تا بتوان از آن به عنوان یک نقطه مرجع استفاده نمود و مقدار واریانس خطاهای برازش شده را با آن مقایسه نمود.

۵.۲ شبیه‌سازی ۲

برای مدل

$$
\begin{aligned}
&x_{i۱} \sim \text{Uniform}(۰, ۱) - \text{Uniform}(۰, ۱), \quad i = ۱, \ldots, n_۱,\\
&x_{i۲} \sim \text{Uniform}(۵, ۳۰) - \text{Uniform}(۵, ۳۰),\\
&x_{i۳} \sim \text{Uniform}(۱۰۰, ۱۵۰),\\
&y_i = ۲ + ۴x_{۱i} + ۶x_{۲i} + x_{۳i} + N(۰, ۱),
\end{aligned}
$$

$$
\begin{aligned}
&x_{i۱} \sim \text{Uniform}(۱۰, ۱۱) - \text{Uniform}(۱۰, ۱۱), \quad i = n_۱ + ۱, \ldots, n,\\
&x_{i۲} \sim \text{Uniform}(۵۰, ۸۰),\\
&x_{i۳} \sim \text{Uniform}(۳۰۰, ۵۰۰),\\
&y_i = ۱ + x_{۱i} + x_{۲i} + x_{۳i} + N(۰, ۱),
\end{aligned}
$$

داده‌ها ۱۰۰۰۰ بار شبیه‌سازی شده‌اند. در هر تکرار مدل‌های کمترین توان‌های دوم خطا، کمترین قدرمطلق انحرافات و مدل پیشنهاد شده در این مقاله برازش شده است. در جهت تحلیل و بررسی عملکرد مدل‌های مورد بحث، خلاصه‌ای از نتایج در جدول ۷ آورده شده است که مانند بخش ۵.۱ تحلیل می‌شوند.

(الف) (ب)

شکل ۳. الف: سد دربند ب: موقعیت جغرافیایی محل گردآوری مشاهدات ناحیه آبخیزداری دربند

۶ تحلیل داده‌های بار معلق آب رودخانه

در مهندسی آب اندازه‌گیری و برآورد صحیح و دقیق بار معلق[1] که توسط رودخانه‌ها حمل می‌شود در بسیاری از پروژه‌های منابع آب بسیار با اهمیت است (چاچی ، ۲۰۱۹). بدین منظور فراهم نمودن نتایج و برآوردهای دقیقی از بار معلق بر مبنای متغیرهای مناسب موجود در ناحیه‌های آبخیزداری[2] یکی از مسایل مهم در مهندسی آب است. در اینباره، جریان سالانه و داده‌های سری زمانی مربوط به روان آب[3] و بار معلق ۱۲ ناحیه آبخیزداری مختلف، شامل سد دربند در رودخانه دربند در خراسان شمالی، ثبت شده است (شکل ۷-الف). سد دربند در رودخانه دربند، در خراسان شمالی و در طول جغرافیایی $''۲۱\ '۳۶\ ^o۳۷$ و عرض جغرافیایی $''۴۸\ '۵۸\ ^o۵۶$ قرار دارد (شکل ۷-ب). مساحت این ناحیه آبخیزداری $km^۲$ ۱۱۱۷ است که بالاترین ارتفاع از سطح دریا آن m ۲۷۲۵ و پایین‌ترین ارتفاع از سطح دریا آن m ۶۸۰ است. لذا بر اساس مطالعات یکسانی در ۱۲ ناحیه آبخیزداری مختلف، با استفاده از ابزارهای استاندارد برخی از خصوصیات آب‌شناسی این نواحی ثبت گردید. در این خصوص از ایستگاه‌های مختلف اندازه‌گیری مشاهدات جدول ۸ مربوط به متغیرهای زیر جمع‌آوری شدند (چاچی ، ۲۰۱۹):

y: میزان بار معلق که توسط رودخانه در یک سال حمل می‌شود ($\frac{ton}{year}$).

[1] Suspended load
[2] Watershed areas
[3] Discharge

$x_۱$: جریان سالیانه آب در رودخانه/ناحیه آبخیزداری ($۱۰^۶ m^۳$).
$x_۲$: مساحت ناحیه آبخیزداری ($km^۲$).
$x_۳$: تجمع سالیانه بار معلق در هر کیلومتر مربع از دلتای رودخانه ($\frac{ton}{year} km^۲$).
$x_۴$: جریان آب مربوط به مقدار تجمع سالیانه بار معلق در هر کیلومتر مربع از دلتای رودخانه ($\frac{۱۰۰۰ m^۳}{km^۲}$).
در سال‌های پر باران که اغلب با سیل همراه است، مشاهداتی به عنوان دورافتاده یا دورافتاده ثبت

جدول ۸. داده‌های مهندسی آب

$x_۴$	$x_۳$	$x_۲$	$x_۱$	y	i
۲۲٫۸۵	۱۵۰	۱۱۱۷	۲۵٫۵	۱۶۳۶۵۶	۱
۹۰٫۷۶	۵۰۴	۴۸۷	۴۴٫۲	۲۴۵۳۲۷	۲
۱۱٫۱۲	۱۷۹	۵۹۲۵	۶۵٫۹	۱۰۶۱۳۵۳	۳
۳۸٫۴۶	۴۰۳	۴۰۳	۱۵٫۵	۱۶۲۳۶۶	۴
۷۶٫۳۳	۲۰۸	۲۰۷	۱۵٫۸	۴۳۱۳۳	۵
۳۸٫۶۹	۳۸۲	۱۲۵۶	۴۸٫۶	۴۸۰۲۷۰	۶
۲۴٫۸۷	۳۰۳	۱۳۵۵	۳۳٫۷	۴۱۰۵۲۴	۷
۴۸٫۳۶	۴۵۲	۵۴۸	۲۶٫۵	۲۴۸۰۰۰	۸
۲۰٫۴۴	۳۲۲	۱۲۰۳۷	۲۴۶٫۰	۳۸۷۷۱۱۲	۹
۱۷٫۳۷	۲۹۱	۱۷۲۶۶	۲۹۹٫۹	۵۰۳۳۴۲۵	۱۰
۱۶٫۳۴	۳۳۲	۲۷۲۴۱	۴۴۵٫۰	۹۰۴۱۴۸۰	۱۱
۶٫۷۸	۶۵۶	۸۸۴	۲۲٫۰	۵۷۹۶۷۰	۱۲

جدول ۹. مدل‌های کاندید برآورد شده بر اساس گام ۳ الگوریتم به همراه برآوردهای LA و LS

$\sum_{i=۱}^{n=۱۲} e_i$	$(a_{۰k}, a_{۱k}, a_{۲k}, a_{۳k}, a_{۴k})$	$\min e_{(k)}$	k
۲۳۲۴۴۰۵	(−۳۰۵۴۲۳٫۰, ۱۲۸۰۵۵٫۵, ۶۷٫۴, ۸۵۶٫۶, −۲۶۸۰٫۴)	$\min e_{(۶)}$	۶
۲۰۱۷۳۰۵	(−۱۹۵۸۸۹٫۰, ۴۶۴۹٫۵, ۲۱۲٫۰, ۷۷۰٫۱, −۲۸۴۹٫۶)	$\min e_{(۷)}$	۷
۱۸۵۳۱۴۱	(−۲۹۹۳۹۳٫۸, ۳۶۶۰٫۴, ۲۷۳٫۷, ۸۶۸٫۲, −۱۹۰۲٫۰)	$\min e_{(۸)}$	۸
۱۹۰۵۹۱۱	(−۲۴۲۴۹۸٫۷, ۱۰۴۶۱٫۴, ۱۱۵٫۲, ۷۹۶٫۴, −۴۷۶۰٫۵)	$\min e_{(۹)}$	۹
۱۷۸۹۲۰۸	(−۲۴۱۲۳۷٫۵, −۲۳۳۶٫۸, ۳۶۹٫۹, ۷۰۹٫۴, ۵۷۴٫۷)	$\min e_{(۱۰)}$	۱۰
۱۹۶۱۵۱۲	(−۲۸۹۸۵۳٫۷, ۱۰۱۶۸٫۶, ۱۱۹٫۹, ۸۴۸٫۶, −۲۴۸۸٫۶)	$\min e_{(۱۱)}$	۱۱
۱۸۲۳۶۴۰	(−۳۷۰۳۴۱٫۸, ۳۳۰۹٫۲, ۲۶۲٫۰, ۷۴۱٫۷, ۲۰۰۰٫۳)	$\min e_{(۱۲)}$	۱۲
۱۶۷۳۷۹۷	(−۳۲۷۸۵۹٫۴, −۲۸۶٫۲, ۳۳۰٫۷, ۹۵۳٫۶, −۵۹۵٫۱)	$\min \sum \lvert e_i \rvert$	LA
۵۶۶۲۵۶۵۱۸۸۴۲	(−۶۵۰۳۱۴٫۸, ۴۲۲۶٫۵, ۲۵۷٫۳, ۱۲۶۷٫۴, ۱۷۳۰٫۶)	$\min \sum e_i^۲$	LS

می‌شوند. لذا در مدل‌سازی این مجموعه مشاهدات بهتر است که از روش‌های استوار در برآورد مدل رگرسیون $y = \alpha_۰ + \alpha_۱ x_۱ + \alpha_۲ x_۲ + \alpha_۳ x_۳ + \alpha_۴ x_۴ + \epsilon$ استفاده شود. با استفاده از رویکرد پیشنهادی، نتایج جداول ۹ و ۱۰ برای مدل‌های مختلف کاندید حاصل شده است. شکل ۲ نمودار میله‌ای مقادیر خطاهای نسبی جدول ۱۰ را نشان می‌دهد. بر طبق این نتایج، در مدل $k = ۱۱$، $\sum_{i=۱}^{۱۱} \delta_i = ۳۴$ درصد از خطای نسبی برازش از یازده مشاهده به دست می‌آید و ۶۶ درصد دیگر از یک مشاهده به دست می‌آید. از اینجا می‌توان نتیجه گرفت که در مجموعه مشاهدات حداقل یک داده دورافتاده وجود دارد.

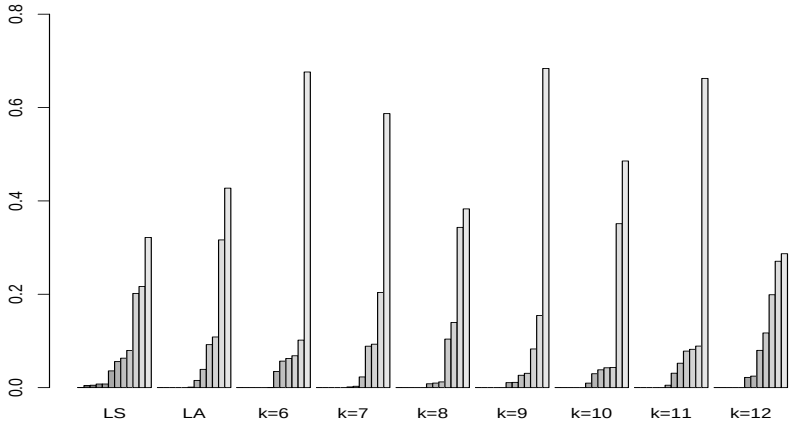
جدول ۱۰. مقادیر خطاهای نسبی مدل‌های کاندید بر اساس گام ۴ الگوریتم در تعیین بهترین مدل بهینه

Δ_k	$\delta_{۱۲}$	$\delta_{۱۱}$	$\delta_{۱۰}$	$\delta_۹$	$\delta_۸$	$\delta_۷$	$\delta_۶$	$\delta_۵$	$\delta_۴$	$\delta_۳$	$\delta_۲$	$\delta_۱$	k
۰٫۰۳۴	۰٫۶۸	۰٫۱۰	۰٫۰۶۸	۰٫۰۶۲	۰٫۰۵۷	۰٫۰۳۴	۰٫۰۰۰	۰	۰	۰	۰	۰	۶
۰٫۰۲۱	۰٫۵۹	۰٫۲۰	۰٫۰۹۳	۰٫۰۸۹	۰٫۰۲۳	۰٫۰۰۳	۰٫۰۰۱	۰	۰	۰	۰	۰	۷
۰٫۰۹۲	۰٫۳۸	۰٫۳۴	۰٫۱۴۰	۰٫۱۰۴	۰٫۰۱۲	۰٫۰۰۹	۰٫۰۰۸	۰	۰	۰	۰	۰	۸
۰٫۰۵۲	۰٫۶۸	۰٫۱۵	۰٫۰۸۳	۰٫۰۳۱	۰٫۰۲۶	۰٫۰۱۱	۰٫۰۱۰	۰	۰	۰	۰	۰	۹
۰٫۳۰۷	۰٫۴۹	۰٫۳۵	۰٫۰۴۳	۰٫۰۴۲	۰٫۰۳۸	۰٫۰۲۹	۰٫۰۰۹	۰	۰	۰	۰	۰	۱۰
۰٫۵۸۰	۰٫۶۶	۰٫۰۸	۰٫۰۸۲	۰٫۰۷۸	۰٫۰۵۲	۰٫۰۳۰	۰٫۰۰۵	۰	۰	۰	۰	۰	۱۱
–	۰٫۲۹	۰٫۲۷	۰٫۱۹۹	۰٫۱۱۷	۰٫۰۸۰	۰٫۰۲۴	۰٫۰۲۱	۰	۰	۰	۰	۰	۱۲
–	۰٫۴۳	۰٫۳۱	۰٫۱۰۸	۰٫۰۹۲	۰٫۰۳۹	۰٫۰۱۵	۰٫۰۰۱	۰	۰	۰	۰	۰	LA
–	۰٫۳۲	۰٫۲۱	۰٫۲۰۱	۰٫۰۷۹	۰٫۰۶۳	۰٫۰۵۵	۰٫۰۳۵	۰٫۰۰۷	۰٫۰۰۷	۰٫۰۰۵	۰٫۰۰۴	۰٫۰۰۰	LS

همچنین تفاوت معنی‌داری بین مقدار $\Delta_{۱۱} = ۰٫۵۸$ برای این مدل و دیگر مدل‌ها وجود دارد. در نتیجه، مدل برازش شده با $k = ۱۱$ به صورت

$$\widehat{y} = -۲۸۹۸۵۳۷ + ۱۰۱۶۸٫۶x_۱ + ۱۱۹۹x_۲ + ۸۴۸٫۶x_۳ - ۲۴۸۸٫۶x_۴,$$

بهترین برازش به این مشاهدات است. بنابراین مشاهده مربوط به $(y_۳, \mathbf{x}_۳)$ یک داده دورافتاده است.



شکل ۴. نمودارهای میله‌ای مقادیر خطاهای نسبی مدل‌های جدول ۱۰

بحث و نتیجه‌گیری

در این مقاله به معرفی و بررسی یک مدل رگرسیون وزنی-مبنا با انحرافات مرتب شده و مقایسه آن با حالات رگرسیون کمترین توان‌های دوم خطا و کمترین قدرمطلق انحرافات پرداخته شد. نتایج مثال‌های عددی به همراه نتایج تحلیل حساسیت برآوردگرهای ارایه شده در مطالعات شبیه‌سازی، نشانگر موثر بودن رویکرد پیشنهادی در تشخیص و بی‌اثرسازی مشاهدات دورافتاده دارد. رویکردی که در این مقاله مورد بحث و تحلیل قرار گرفت، جایگزین مناسبی برای مدل‌های رگرسیون مبتنی بر رویکرهای رگرسیون کمترین توان‌های دوم خطا و کمترین قدرمطلق انحرافات و حتی M-برآوردگرها است. بر طبق بردار وزن اختیار شده، حالات متنوعی از مدل‌های رگرسیون متداول از قبیل رگرسیون کمترین توان‌های دوم خطا و کمترین قدرمطلق انحرافات و حتی مدل‌های رگرسیون پایدار از قبیل LTS، LMS و رگرسیون چندکی همراه با نقاط شکست بالا حاصل می‌شود. لذا کلاسی از فرمول‌ها و مدل‌های و مسایل بهینه‌سازی به وجود می‌آورد که روش‌های رگرسیون پایدار متداول را پوشش می‌دهد. در این کلاس همچنین می‌توان بر مبنای بردارهای وزن اختیار شده، الگوریتم‌ها و روش‌های حل عددی مختلفی معرفی نمود. در این مقاله روش بهینه‌سازی خاصی در کمینه‌سازی تابع برازش مورد استفاده قرار گرفت. این روش به درستی و با دقت قادر به شناسایی و کم اثر ساختن نقاط دورافتاده است. همچنین، بررسی نظری خواص برآوردگرهای ارایه شده به همراه بررسی خواص حدی چنین برآوردگرهایی می‌تواند در کارهای آتی مورد بررسی و تحقیق قرار گیرد.

تقدیر و تشکر

نویسندگان مقاله از داوران گرامی و ویراستار محترم مجله که نظرات ارزشمند ایشان باعث بهبود مطالب ارایه شده در این مقاله گردید، کمال تشکر و قدردانی را دارند. مقاله حاضر برگرفته از طرح پژوهشی به شماره $IR/۲۱.MA/۱۰۲$ در دانشگاه صنعتی شهدای هویزه است. نویسنده اول از اعتبارات پژوهشی به شماره $SCU.MS/۹۸/۳۸۸۳۷$ در دانشگاه شهید چمران اهواز و نویسنده دوم از اعتبارات پژوهشی به شماره ۹۷/۲۰۰/۱۰۷۳ در دانشگاه صنعتی شهدای هویزه استفاده نموده است.

مراجع

محمودی، م. (۱۳۸۴). رگرسیون حداقل قدرمطلق انحرافات وزنی استوار، پایان‌نامه کارشناسی ارشد، دانشکده ریاضی و کامپیوتر، دانشگاه صنعتی امیرکبیر.

Atkinson, A. and Riani. M. (2000), *Robust Diagnostic Regression Analysis*, Springer, New York.

Becker, C. and Gather, U. (1999), The Masking Breakdown Point of Multivariate Outlier Identification Rules, *Journal of the American Statistical Association*, **94**, 947-55.

Billor, N., Chatterjee, S. and Hadi, A. S. (2006), A Re-weighted Least Squares Method for Robust Regression Estimation, *American Journal of Mathematical and Management Sciences*, **26**, 229-52.

Chachi, J. (2019), A Weighted Least Squares Fuzzy Regression for Crisp Input Fuzzy Output Data, *IEEE Transactions on Fuzzy Systems*, **27**, 739-748.

Chaji, A. R. (2017), Analytic Approach on Maximum Bayesian Entropy Ordered Weighted Averaging Operators, *Computers and Industrial Engineering*, **105**, 260-264.

Chaji, A. R., Fukuyama, H. and Shiraz, R. K. (2018), Selecting a Model for Generating OWA Operator Weights in MGADM Problems by Maximum Entropy Membership Function, *Computers and Industrial Engineering*, **124**, 370-378.

Dervilis, N., Worden, K. and Cross, E. J. (2015), On Robust Regression Analysis as a Mean of Exploring Environmental and Operational Conditions for SHM Data. *Journal of Sound and Vibration*, **347**, 279-296.

Gao, X. and Feng, Y. (2018), Penalized Weighted Least Absolute Dviation Regression, *Statatistical Interface*, **11**, 79-89.

Hawkins, D. M. (1980), *Identification of Outliers*, Chapman and Hall, London.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984), Location of Several Outliers in Multiple Regression Using Elemental Sets, *Technometrics*, **26**, 197-208.

Huber, P. and Ronchetti, E. M. (2009), *Robust Statistics*, 2ed., Wiley, Hoboken, NJ.

Hubert, M., Rousseeuw, P. J. and VanAelst, S. (2008), High-breakdown Robust Multivariate Methods, *Statistical Science*, **23**, 92-119.

Kordos, M., Arnaiz-González, Á. and García-Osorio, C. (2019), Evolutionary Prototype Selection for Multi-Output Regression, *Neurocomputing* **358**, 309-320.

Leite, D. and Škrjanc, I. (2019), Ensemble of Evolving Optimal Granular Experts, OWA Aggregation, and Time Series Prediction, *Information Sciences*, **504**, 95-112.

Ling, S. (2005), Self-Weighted Least Absolute Deviation Estimation for Infinite Variance Autoregressive Models, *Journal of the Royal Statistical Society, Series B*, **67**, 381-393.

Marubini, E. and A. Orenti. (2014), Detecting Outliers and/or Leverage Points: A Robust Two-Stage Procedure with Bootstrap Cut-Off Points, *Epidemiology, Biostatistics and Public Health*, **11**, 90-94.

Nguyen, T. D. and Welsch, R. (2010), Outlier Detection and Least Trimmed Squares Approximation Using Semi-Definite Programming, *Computational Statistics and Data Analysis*, **54**, 3212-3226.

Ogundele, S. O., Mbegbu, J. I. and Nwosu, C. R. (2016), An Alternative Algorithm and R Programming Implementation for Least Absolute Deviation Estimator of the Linear Regression Models, *Journal of Modern Applied Statistical Methods*, **15**, 755-767.

Rousseeuw, P. J. (1984), Least Median of Squares Regression, *Journal of the American Statistical Association*, **79**, 871-80.

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

Rousseeuw, P. J. and Hubert, M. (2013), High-Breakdown Estimators of Multivariat Elocation and Scatter, In *Robustness and Complex Data Structures*, (Eds C. Becker, R. Fried and S. Kuhnt), Berlin: Springer, 49-66.

Rousseeuw, P. J. and VanZomeren, B. C. (1990), Unmasking Multivariate Outliers and Leverage Points, *Journal of the American Statistical Association*, **85**, 633-639.

Salini, S., Cerioli, A., Laurini, F. and Riani, M. (2016), Reliable Robust Regression Diagnostics, *International Statistical Review*, **84**, 99-127.

Vic, B. and Lewis, T. (1994), *Outliers in Statistical Data*, Wiley, NewYork.

Yari, G. and Chaji, A. R. (2012a), Maximum Bayesian Entropy Method for Determining Ordered Wighted Averaging Operator Weights, *Computers and Industrial Engineering*, **63**, 338-342.

Yari, G. and Chaji, A. R. (2012b), Determination of Ordered Weighted Averaging Operator Weights Based on the M-Entropy Measures, *International Journal of Intelligent Systems*, **27**, 1020-1033.

Journal of Statistical Sciences, Spring and Summer, 2021
Vol. 15, No. 1, pp 39-60
DOI: 10.29252/jss.15.1.57

Employing Weighted Operators in Ordered Least Deviations Regression Model

Chachi, J.[1], Chaji, A.[2]
[1]Department of Statistics, Shahid Chamran University of Ahvaz, Ahvaz, Iran.
[2]Department of Electrical Engineering, Shohadaye Hoveizeh University of Technology, Dashte-Azadegan, Iran.

Abstract: This article introduces a new method to estimate the least absolutes linear regression model's parameters, which considers optimization problems based on the weighted aggregation operators of ordered least absolute deviations. In the optimization problem, weighted aggregation of orderd fitted least absolute deviations provides data analysis to identify the outliers while considering different fitting functions simultaneously in the modeling problem. Accordingly, this approach is not affected by outlier observations and in any problem proportional to the number of potential outliers selects the best model estimator with the optimal break-down point among a set of other candidate estimators. The performance and the goodness-of-fit of the proposed approach are investigated, analyzed and compared in modeling analytical dataset and a real value dataset in hydrology engineering at the presence of outliers. Based on the results of the sensitivity analysis, the properties of unbiasedness and efficiency of the estimators are obtained.

Keywords: Weighted regression, Ordered absolutes-deviations, Breakdown point, Squared errors.

Mathematics Subject Classification (2010): 62J05, 62J02, 62G08.