برآوردگر استوار مرزبندی شده تعمیم‌یافته محتمل در مدل رگرسیون نیمه‌پارامتری

مهدی روزبه[۱]، مرتضی امینی[۲]

[۱]گروه آمار، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه سمنان

[۲]گروه آمار، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران

چکیده: در تجزیه و تحلیل مسائل رگرسیونی و به‌ویژه مدل بندی آماری بسیاری از داده‌ها مانند داده‌های اقتصادی، روانشناسی، علوم اجتماعی، علوم پزشکی، مهندسی و غیره با مشکل هم‌خطی در میان متغیرهای پیشگو و حضور نقاط دورافتاده در مجموعه داده‌ها مواجه می‌شویم. در چنین مواقعی برآوردگر کمترین توان‌های دوم معمولی منجر به برآوردگرهای نادقیق می‌شود. برای غلبه بر مشکل مشاهده‌های دورافتاده از روش‌های استوار استفاده می‌شود. همچنین برای حل مشکل هم‌خطی چندگانه استفاده از رگرسیون مرزبندی شده توصیه می‌شود. از طرف دیگر در شرایطی که واریانس خطاها ناهمگن بوده یا خطاها دارای خودهمبستگی باشند، از روش کم‌ترین توان‌های دوم تعمیم‌یافته استفاده می‌شود. در این مقاله ابتدا یک الگوریتم سریع برای محاسبه برآوردگر کم‌ترین توان‌های دوم تعمیم‌یافته پیراسته مرزبندی شده محتمل در مدل رگرسیون نیمه‌پارامتری پیشنهاد شده و سپس با استفاده از شبیه‌سازی به روش مونت کارلو و یک داده واقعی، کارایی برآوردگرهای پیشنهادی سنجیده می‌شود.

واژه‌های کلیدی: اعتبارسنجی متقابل، برآوردگر کم‌ترین توان‌های دوم پیراسته، مدل رگرسیون نیمه‌پارامتری، نقاط دورافتاده، نقطه شکست.

°آدرس الکترونیک مسئول مقاله: مهدی روزبه، mahdi.roozbeh@semnan.ac.ir

°کد موضوع‌بندی ریاضی (۲۰۱۰): 62G08، 62G35، 62J20.

۱ مقدمه

مدل رگرسیون نیمه‌پارامتری یکی از پرکاربردترین مدل‌های رگرسیون در حوزه یادگیری آماری است که در سال‌های اخیر مورد توجه محققین قرار گرفته است. علاوه بر کاربرد گسترده این مدل‌ها در تحلیل کواریانس، آن‌ها کاربرد زیادی در مسائل اقتصادی دارند که از مهمترین آن‌ها می‌توان به مدل تابع درآمد سرمایه انسانی (ویلیس، ۱۹۸۶) و منحنی دستمزد (بلنچفلاور و اسوالد، ۱۹۹۴) اشاره کرد. در هر دو مورد، لگاریتم درآمد شخصی به ویژگی‌های شخصیتی (مانند جنسیت و وضعیت تاهل) و اندازه سرمایه‌های انسانی فرد (مانند تحصیل و تجربه کار در بازار) وابسته است. در نظریه اقتصاد رابطه غیرخطی بین لگاریتم درآمد شخصی و تجربه کار در بازار مورد بررسی قرار می‌گیرد. این در حالی است که رابطه امیدریاضی لگاریتم درآمد شخصی و متغیرهای کیفی جنسیت، وضعیت تأهل و تحصیلات خطی است.

فرض کنید مشاهدات مستقل $(\boldsymbol{x}_n, t_n, y_n)$،...،$(\boldsymbol{x}_1, t_1, y_1)$ در مدل

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + f(\boldsymbol{t}_i) + \boldsymbol{\epsilon}_i,\ i = 1, \ldots, n, \qquad (1)$$

صدق کنند، که در آن y_iها مشاهدات تک‌بعدی متغیر پاسخ، $\boldsymbol{x}_i$ها بردارهای pبعدی متغیرهای پیشگو دارای رابطه خطی با متغیر پاسخ، $\boldsymbol{\beta}$ بردار pبعدی ضرایب نامعلوم، t_iها مشاهدات متغیر یا متغیرهای پیشگو دارای رابطه غیرخطی با متغیر پاسخ، $f(\cdot)$ تابعی نامعلوم و هموار با دامنه $D \subset \mathcal{R}$ و ϵ_iها خطاهای تصادفی هستند. در این مدل فرض می‌شود که تابع نامعلوم $f(\cdot)$ هموار بوده و همان‌طور که ملاحظه می‌شود، مدل‌های رگرسیونی نیمه‌پارامتری انعطاف‌پذیرتر از مدل‌های رگرسیون خطی هستند، زیرا هم دارای قسمت خطی و هم قسمت غیرخطی می‌باشند. در واقع استفاده از این مدل زمانی مفید است که متغیر وابسته y به‌طور خطی با متغیر مستقل $\boldsymbol{x}$ و به‌طور غیرخطی با متغیر مستقل t رابطه داشته باشد. محققان این مدل را به مدل ناپارامتری محض و رگرسیون خطی ترجیح می‌دهند. مدل‌های رگرسیون نیمه‌پارامتری نخستین بار توسط انگل و همکارانش در سال ۱۹۸۶ در بررسی رابطه میان مصرف ماهیانه برق به‌عنوان متغیر پاسخ و قیمت ماهیانه برق، درآمد ماهیانه و دمای هوا به‌عنوان متغیرهای پیشگو با در نظر گرفتن قیمت ماهیانه برق و درآمد ماهیانه به‌عنوان بخش خطی و دمای هوا به‌عنوان بخش غیرخطی مورد استفاده قرار گرفتند.

هم‌خطی چندگانه یکی از مشکلات متداول در مدل‌های رگرسیون خطی چندگانه است که به دلیل ایجاد تورم در ماتریس کواریانس، باعث کاهش کارایی برآوردگرهای کلاسیک و نهایتا پیش‌بینی نادقیق می‌شود. رگرسیون

تاوانیده[1] یک راهکار منظم‌سازی است که برآوردگرها را با اضافه کردن یک جمله تاوان به تابع هدف بهبود می‌بخشد. این روش تورم (بزرگی مقدارهای) ماتریس کواریانس برآورد ضرایب را با انقباض برآوردگرها به سمت صفر کاهش می‌دهد. برآوردگر مرزبندی شده[2] (هورل و کنارد، ۱۹۷۰) در واقع برآوردگر رگرسیون تاوانیده با تابع تاوان توان دوم نرم L_2 ضرایب رگرسیونی است. این روش یکی از پرکاربردترین راهکارها در مقابله با مشکل هم‌خطی چندگانه است. خصوصیات نظری برآوردگر مرزبندی شده به‌طور گسترده‌ای مورد مطالعه قرار گرفته است. به‌عنوان مثال می‌توان به فاربرادر (۱۹۷۶)، گلاب و همکاران (۱۹۷۹)، اسپکمن (۱۹۸۸)، آرشی و همکاران (۲۰۱۵) و نوروزی‌راد و آرشی (۱۳۹۶) اشاره کرد. همچنین، این برآوردگر در مدل رگرسیون نیمه‌پارامتری توسط تاباکان و آکدنیز (۲۰۱۰)، آکدنیز و همکاران (۲۰۱۲)، روزبه (۲۰۱۵)، آرشی و ولی‌زاده (۲۰۱۵) و امینی و روزبه (۲۰۱۵)، مورد مطالعه قرار گرفته است. با این حال برآوردگر کلاسیک مرزبندی شده مانند برآوردگر کلاسیک کم‌ترین توان‌های دوم نسبت به مشاهدات دورافتاده حساس است. مشاهدات دورافتاده نقاطی هستند که از مدل اکثریت نقاط تبعیت نمی‌کنند. جانشین‌های استوار روش کلاسیک کمترین توان‌های دوم مانند M-برآوردگرها، S-برآوردگرها، برآوردگر حداقل میانه توان‌های دوم $(LMS$ و روش کم‌ترین توان‌های دوم پیراسته[3] (LTS) برای کاهش اثر مشاهدات دورافتاده مورد استفاده قرار گرفته‌اند. برای مروری بر این روش‌ها می‌توان به مارونا و همکاران (۲۰۰۶) مراجعه کرد. روش کم‌ترین توان‌های دوم پیراسته که نخستین بار توسط روسیو (۱۹۸۴) معرفی شد، به دلیل تعریف ساده، نقطه فروریزش بالا (بخش ۴ را ببینید) و استواری در برابر نقاط نافذ (مشاهدات دور افتاده در جهت متغیرهای پیشگو) یکی از متداول‌ترین روش‌های برآورد استوار است. برای یک مجموعه شامل n مشاهده و یک پارامتر صحیح $h \leq n$، برآوردگر LTS ابرصفحه حداقل کننده مجموع کوچکترین h مربع مانده را به مشاهدات برازش می‌دهد. اگر یکی از متغیرهای پیش‌بین یک تابع دقیق خطی از یک یا چند متغیر پیش‌بین دیگر باشد، مدل رگرسیون دارای هم‌خطی کامل است. هم‌خطی ناقص وقتی اتفاق می‌افتد که یکی از متغیرهای پیش‌بین به‌طور تقریبی یک تابع خطی از یک یا چند متغیر پیش‌بین دیگر باشد. برآوردگرهای استوار کمترین توان‌های دوم پیراسته در مدل رگرسیون توسط روزبه و بابایی کفاکی (۲۰۱۶) و در مدل رگرسیون نیمه‌پارامتری توسط تورکمن و تاباکان (۲۰۱۵) مطالعه شده است. آرشی و نوروزی‌راد (۲۰۱۵) M-برآوردگرهای مرزبندی شده بهبودیافته را در مدل رگرسیون تعمیم

[1] Penalized
[2] Ridge
[3] Least Treamed Squares

دادند. در این مقاله، به معرفی حالت تعمیم‌یافته برآوردگرهای مرزبندی شده استوار محتمل[۴] در مدل رگرسیون نیمه‌پارامتری برای استفاده در حالت ناهمگنی واریانس خطاها و مدل رگرسیون سری زمانی پرداخته می‌شود.

در بخش ۲ مقدمات نظری و فرضیات اساسی بیان می‌شوند. بخش ۳ به معرفی برآوردگرهای پارامتری و نیمه‌پارامتری ضرایب و تابع f اختصاص داده شده است. نقطه فروریزش برآوردگر، برآورد پارامترهای اریب‌سازی و پهنای باند به ترتیب مطالب بخش‌های ۴ و ۵ را تشکیل می‌دهند. در پایان نتایج عددی شامل بررسی شبیه‌سازی به روش مونت کارلو و تحلیل داده‌های واقعی در بخش ۶ ارائه شده و در انتها به بحث و نتیجه‌گیری پرداخته می‌شود.

۲ فرض‌های اساسی

شکل ماتریسی مدل رگرسیون نیمه‌پارامتری به‌صورت

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{۲}$$

تعریف می‌شود، که در آن $\boldsymbol{y} = (y_1, \ldots, y_n)'$ بردار $n \times 1$ از مشاهدات پاسخ، $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ ماتریس $n \times p$ متغیرهای پیشگوی بخش خطی، $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ بردار $p \times 1$ ضرایب نامعلوم بخش خطی، $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))'$ بردار $n \times 1$ اثر غیرخطی و $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ بردار $n \times 1$ خطای تصادفی هستند. در این مدل فرض ناهمگنی واریانس خطاها یا ناهمبستگی خطاها به صورت

$$E(\boldsymbol{\epsilon}|\boldsymbol{X}, \boldsymbol{t}) = \boldsymbol{o}\ , Var(\boldsymbol{\epsilon}|\boldsymbol{X}, \boldsymbol{t}) = \sigma^2 \boldsymbol{V},$$

نوشته می‌شود، که در آن σ پارامتری مجهول و $\boldsymbol{V}$ یک ماتریس معین مثبت نامعلوم است. با فرض معلوم بودن $\boldsymbol{\beta}$، می‌توان تابع $f(\cdot)$ را با استفاده از روش هموارسازی تابع رگرسیونی پاسخ $y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ بر روی پیشگوی t_i به‌صورت

$$\hat{f}(t) = \sum_{i=1}^{n} k_i(t)(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \tag{۳}$$

برآورد کرد، به‌طوری که k_i توابع وزنی مثبت بوده و به مشاهدات بستگی دارند.

[4] Feasible

در این مقاله از برآوردگر کرنل گاسر–مولر استفاده می‌شود، که در آن وزن‌ها به صورت

$$k_i(t) = \frac{۱}{\omega}\int_{s_{i-۱}}^{s_i} K\big(\frac{t-s}{\omega}\big)\, ds\ , s_i = (t_i + t_{i+۱})/۲\ , s_۰ = ۰\ , s_۱ = ۱$$

تعریف می‌شوند (برای مطالعه بیشتر و خصوصیات این برآوردگر می‌توان به یوبانگ ۱۹۹۹)) مراجعه کرد)، به‌طوری‌که $K(\cdot)$ تابع هسته و ω پارامتر پهنای باند هستند. همچنین فرض می‌شود که $K(\cdot)$ یک تابع هسته رتبه دوم ($K'' \neq ۰$) با خصوصیات

$$\int_{-۱}^{۱} K(u)\,\mathrm{d}u = ۱,\ \int_{-۱}^{۱} uK(u)\,\mathrm{d}u = ۰,$$
$$M_۲ = \int_{-۱}^{۱} u^۲K(u)\,\mathrm{d}u \neq ۰,\ \kappa_۲ = \int_{-۱}^{۱} K(u)^۲\,\mathrm{d}u < \infty,$$

باشد. تحت شرایط بالا بنابر یوبانگ (۱۹۹۹) می‌توان نوشت

$$\mathrm{E}\big(\hat{f}(t) - f(t)\big) = \frac{\omega^۲ f''(t) M_۲}{۲} + O(\omega^۲) + O(n^{-۱})$$
$$\mathrm{Var}\big(\hat{f}(t)\big) = (n\omega)^{-۱}\sigma^۲\kappa_۲ + O((n\omega)^{-۲}).$$

به منظور برآورد بردار پارامتر خطی $\boldsymbol{\beta}$ به‌روش مانده‌های جزئی، با جایگذاری مقدار $\hat{f}(\cdot)$ در معادله (۱)، مدل

$$y_i - \sum_{j=۱}^{n} k_j(t_i)y_j = \big(\boldsymbol{x}_i' - \sum_{j=۱}^{n} k_j(t_i)\boldsymbol{x}_j'\big)\boldsymbol{\beta} + \epsilon_i,$$

حاصل می‌شود. اکنون با در نظر گرفتن $\breve{y}_i = y_i - \sum_{j=۱}^{n} k_j(t_i)y_j$ و $\breve{\boldsymbol{x}}_i' = \boldsymbol{x}_i' - \sum_{j=۱}^{n} k_j(t_i)\boldsymbol{x}_j'$ می‌توان مدل اخیر را به صورت $\breve{y}_i = \breve{\boldsymbol{x}}_i'\boldsymbol{\beta} + \epsilon_i$ نوشت، که همان مدل رگرسیون استاندارد است. بنابراین، با تعریف $\breve{\boldsymbol{X}} = (\boldsymbol{I}_n - \boldsymbol{K})\boldsymbol{X}$ و $\breve{\boldsymbol{y}} = (\boldsymbol{I}_n - \boldsymbol{K})\boldsymbol{y}$ که $\boldsymbol{K}$ ماتریس هموارساری در (i,j)امین مولفه برابر $k_j(t_i)$ برای $i, j = ۱, \ldots, n$ است، برآورد کلاسیک کم‌ترین توان های دوم تعمیم‌یافته $\boldsymbol{\beta}$ در مدل جدید $\breve{\boldsymbol{y}} = \breve{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ به‌صورت

$$\hat{\boldsymbol{\beta}}_G = \big(\breve{\boldsymbol{X}}'\boldsymbol{V}^{-۱}\breve{\boldsymbol{X}}\big)^{-۱}\breve{\boldsymbol{X}}'\boldsymbol{V}^{-۱}\breve{\boldsymbol{y}}, \qquad (۴)$$

خواهد بود، که در آن $\breve{\boldsymbol{y}} = (\breve{y}_۱, \ldots, \breve{y}_n)'$ بردار پاسخ هموار شده و $\breve{\boldsymbol{X}} = (\breve{\boldsymbol{x}}_۱, \ldots, \breve{\boldsymbol{x}}_n)'$ ماتریس طرح هموار شده است. جائو و همکاران (۱۹۹۵) ثابت کردند که برآوردگر به‌دست آمده برای $\boldsymbol{\beta}$، برآوردگری سازگار و دارای توزیع نرمال مجانبی به صورت

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}) \xrightarrow{D} N\big(۰, \sigma_\epsilon^۲\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{t}}^{-۱}\big),$$

است. در عمل، برای استنباط راجع به پارامتر $\boldsymbol{\beta}$، باید σ_ϵ^2 و $\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{t}}$ برآورد شوند. برآوردگرهای سازگار

$$s_\epsilon^2 = \frac{1}{n}(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_G)' \boldsymbol{V}^{-1} \left(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_G\right) \xrightarrow{p} \sigma_\epsilon^2$$
$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}|t} = \frac{1}{n}(\check{\boldsymbol{X}}' \boldsymbol{V}^{-1} \check{\boldsymbol{X}}) \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{x}|t}.$$

رهیافتی برای حل این مساله می‌باشند. در مرحله آخر، می توان تابع $f(\cdot)$ را با جایگذاری برآوردگر نیمه‌پارامتری سازگار به‌دست آمده برای $\boldsymbol{\beta}$ در معادله (۳)، با استفاده از برآوردگر

$$\hat{f}(t) = \boldsymbol{k}(t)(\boldsymbol{y} - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_G), \tag{۵}$$

برآورد کرد. بنابر هاردل و همکاران (۲۰۰۰) برآوردگر $\hat{f}(t)$ یك برآوردگر سازگار برای $f(t)$ است.

۳ برآورد مرزبندی شده کم‌ترین توان‌های دوم تعمیم‌یافته پیراسته محتمل

با ضرب ماتریس $\boldsymbol{V}^{-1/2}$ در طرفین معادله $\check{\boldsymbol{y}} = \check{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ به منظور ناهمبسته نمودن جملات خطا و ثابت نمودن واریانس آن‌ها، می‌توان مدل جدید

$$\boldsymbol{V}^{-1/2}\check{\boldsymbol{y}} = \boldsymbol{V}^{-1/2}\check{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{V}^{-1/2}\boldsymbol{\epsilon},$$

را در نظر گرفت. حال با توجه به $Var(\boldsymbol{V}^{-1/2}\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$، می‌توان مانده‌های تبدیل‌یافته $\mathbf{e}(\boldsymbol{\beta}) = \boldsymbol{V}^{-1/2}(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{\beta})$ را با مانده‌های $(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{\beta})$ جایگزین کرد. در این صورت برآوردگر استوار کم‌ترین توان‌های دوم پیراسته تعمیم‌یافته به‌صورت

$$\hat{\boldsymbol{\beta}}_{GLTS}(\boldsymbol{z}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{h} (\mathbf{e}^2(\boldsymbol{\beta}))_{i:n} = \underset{\boldsymbol{\beta}, \mathbf{z}\in E_h}{\operatorname{argmin}} (\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{\beta})' \boldsymbol{V}^{-1/2}\mathbf{Z}\boldsymbol{V}^{-1/2}(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{\beta}),$$

تعریف می‌شود، به‌طوری‌که $(\mathbf{e}^2(\boldsymbol{\beta}))_{1:n} \le \cdots \le (\mathbf{e}^2(\boldsymbol{\beta}))_{n:n}$ آماره‌های مرتب توان دوم مانده‌ها و $\boldsymbol{Z}$ ماتریس قطری با عناصر قطر اصلی $\boldsymbol{z} = (z_1, \ldots, z_n)'$ است. اگر مشاهده iام خوب باشد، z_i مقدار یک و اگر مشاهده iام پرت باشد، z_i مقدار صفر را اختیار می‌کند. همچنین

$$E_h = \left\{\boldsymbol{z};\ z_i \in \{0, 1\},\ i = 1, \ldots, n, \boldsymbol{z}'\mathbf{1} = h\right\},\ h \le n.$$

در حالت کلی $h = [[n(۱-\alpha)]]$ در نظر گرفته می‌شود که $\alpha \in (۰, ۱)$ نسبت مشاهدات تاثیرگذار و $[[x]]$ مقدار صحیح عدد اعشاری x است. در واقع $۱ - \alpha$ یک حدس اولیه برای درصد مشاهدات غیر دورافتاده است که توسط برخی از محققان ۰/۷۵ در نظر گرفته شده است. با چنین انتخابی برآورد نهایی بر اساس درصد قابل توجهی از مشاهدات به دست آمده و نقطه فروریزش برآوردگر نیز ۲۵٪ خواهد بود. برخی دیگر از محققان مقدار h را به‌صورت $h = [[n/۲]] + [[(p+۱)/۲]]$ پیشنهاد داده‌اند (آلفونس و همکاران، ۲۰۱۳).

در حضور هم‌خطی چندگانه و مشاهدات دورافتاده، می‌توان از برآوردگر کم‌ترین توان‌های دوم پیراسته مرزبندی شده تعمیم‌یافته به‌صورت

$$\hat{\boldsymbol{\beta}}_{GLTS}(\boldsymbol{z}, \lambda) = \underset{\boldsymbol{\beta}, \mathbf{z} \in E_h}{\operatorname{argmin}} Q(\boldsymbol{z}, \boldsymbol{\beta}),$$

استفاده کرد، که در آن $Q(\boldsymbol{z}, \boldsymbol{\beta}) = (\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{\beta})' \boldsymbol{V}^{-۱/۲} \boldsymbol{Z} \boldsymbol{V}^{-۱/۲} (\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}'\boldsymbol{\beta}$. با قرار دادن $\boldsymbol{Z}^* = \operatorname{Diag}(\boldsymbol{z}^*)$ که در آن

$$\boldsymbol{z}^* = \operatorname{argmin}_{E_h} \big(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{\beta}(\boldsymbol{z}, \lambda)\big)' \boldsymbol{V}^{-۱/۲} \boldsymbol{Z} \boldsymbol{V}^{-۱/۲} \big(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{\beta}(\boldsymbol{z}, \lambda)\big),$$
$$\boldsymbol{\beta}(\boldsymbol{z}, \lambda) = (\check{\boldsymbol{X}}' \boldsymbol{V}^{-۱/۲} \boldsymbol{Z} \boldsymbol{V}^{-۱/۲} \check{\boldsymbol{X}} + \lambda \boldsymbol{I}_p)^{-۱} \check{\boldsymbol{X}}' \boldsymbol{V}^{-۱/۲} \boldsymbol{Z} \boldsymbol{V}^{-۱/۲} \check{\boldsymbol{y}},$$

برآوردگر کم‌ترین توان‌های دوم پیراسته مرزبندی شده تعمیم‌یافته برابر است با

$$\hat{\boldsymbol{\beta}}_{GLTS}(\boldsymbol{z}^*, \lambda) = (\check{\boldsymbol{X}}' \boldsymbol{V}^{-۱/۲} \boldsymbol{Z}^* \boldsymbol{V}^{-۱/۲} \check{\boldsymbol{X}} + \lambda \boldsymbol{I}_p)^{-۱} \check{\boldsymbol{X}}' \boldsymbol{V}^{-۱/۲} \boldsymbol{Z}^* \boldsymbol{V}^{-۱/۲} \check{\boldsymbol{y}}. \quad (۶)$$

همچنین برآورد استوار تابع $f(\cdot)$ در هر نقطه $t \in D$ با استفاده از رابطه

$$\hat{\boldsymbol{f}}_{GLTS}(\boldsymbol{t}; \boldsymbol{z}^*, \lambda) = \boldsymbol{k}(t)\big(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLTS}(\boldsymbol{z}^*, \lambda)\big), \quad (۷)$$

محاسبه می‌شود. با توجه به‌صورت برآوردگرهای اخیر (در عین کاربردی بودن آن)، تنها در حالاتی که ماتریس $\boldsymbol{V}$ معلوم است می‌توان نتایجی دقیق به‌دست آورد، ولی این حالت مسائل کاربردی و واقعی را پوشش نمی‌دهد. به‌منظور کاربردی بودن و استفاده از مشاهدات واقعی تنها می‌توان رفتار مجانبی برآوردگرهای اخیر را مورد تجزیه و تحلیل قرار داد. برای مشاهده توجیه و دلیل این مطلب می‌توان به زلنر (۱۹۶۲ و۱۹۶۳) مراجعه کرد.

چون ماتریس کوواریانس خطا یعنی $\boldsymbol{V}$ در عمل معمولا نامعلوم است، بنابراین استفاده از $\hat{\boldsymbol{\beta}}_{GLTS}(\boldsymbol{z}^*, \lambda)$ غیرممکن بوده و باید ماتریس $\boldsymbol{V}$ بوسیله یك برآوردگر مناسب جایگزین شود. زلنر (۱۹۶۲) ثابت کرد

$\boldsymbol{S} = \frac{۱}{n-p)}(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{b})(\check{\boldsymbol{y}} - \check{\boldsymbol{X}}\boldsymbol{b})'$ یک برآوردگر سازگار برای $\boldsymbol{V}$ است، به‌طوریکه $\boldsymbol{b}$ برآوردگر کم‌ترین توان‌های دوم معمولی $\boldsymbol{\beta}$ یعنی $(\check{\boldsymbol{X}}'\check{\boldsymbol{X}})^{-۱}\check{\boldsymbol{X}}'\check{\boldsymbol{y}}$ است. با جایگزینی $\boldsymbol{S}$ در عبارت (۴) بجای $\boldsymbol{V}$، برآوردگر کم‌ترین توان‌های دوم تعمیم‌یافته محتمل به صورت

$$\hat{\boldsymbol{\beta}}_{FG} = (\check{\boldsymbol{X}}'\boldsymbol{S}^{-۱}\check{\boldsymbol{X}})^{-۱}\check{\boldsymbol{X}}'\boldsymbol{S}^{-۱}\check{\boldsymbol{y}},$$

به‌دست می‌آید. بنابراین، بر اساس برآوردگر پیشنهادی اخیر توسط زلنر (۱۹۶۲)، می‌توان برآوردگر کم‌ترین توان‌های دوم پیراسته مرزبندی شده تعمیم‌یافته محتمل

$$\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*, \lambda) = (\check{\boldsymbol{X}}'\boldsymbol{S}^{-۱/۲}\boldsymbol{Z}^*\boldsymbol{S}^{-۱/۲}\check{\boldsymbol{X}} + \lambda\boldsymbol{I}_p)^{-۱}\check{\boldsymbol{X}}'\boldsymbol{S}^{-۱/۲}\boldsymbol{Z}^*\boldsymbol{S}^{-۱/۲}\check{\boldsymbol{y}}, \quad (۸)$$

را با جایگزینی $\boldsymbol{S}$ در عبارت (۹) بجای $\boldsymbol{V}$ به‌دست آورد. با استفاده از زلنر (۱۹۶۲) می‌توان نتیجه گرفت که $\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*, \lambda) = \hat{\boldsymbol{\beta}}_{GLTS}(\boldsymbol{z}^*, \lambda) + O(n^{-۱})$ و بنابراین، $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*, \lambda) - \boldsymbol{\beta}\big)$ و $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{GLTS}(\boldsymbol{z}^*, \lambda) - \boldsymbol{\beta}\big)$ دارای توزیع مجانبی نرمال یکسان هستند. بنابراین، هنگامی که ماتریس کوواریانس جملات خطا نامعلوم است، برآورد استوار محتمل تابع $f(\cdot)$ در هر نقطه $t \in D$ با استفاده از

$$\hat{\boldsymbol{f}}_{FGLTS}(\boldsymbol{t}; \boldsymbol{z}^*, \lambda) = \boldsymbol{k}(t)\big(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*, \lambda)\big), \quad (۹)$$

محاسبه می‌شود.

۴ نقطه فروریزش برآوردگر استوار مرزبندی شده محتمل

یک معیار برای میزان استواری یک برآوردگر، نقطه فروریزش نمونه متناهی است.

تعریف ۱. (مارونا و همکاران، ۲۰۰۶) نقطه فروریزش برآوردگر $T = T(\mathbf{Z})$ برای نمونه $\mathbf{Z}$ به حجم n به‌صورت

$$\mathrm{BP}(T; \mathbf{Z}) = \min_{m} \big\{\frac{m}{n} : \sup_{\mathbf{Z}^*} ||T(\mathbf{Z}^*)||^{۲} = \infty\big\},$$

تعریف می‌شود، که در آن $\mathbf{Z}^*$ نمونه آلوده شده[۵] با جایگزینی $m \leq n$ نقطه از نمونه $\mathbf{Z}$ با مقادیر دلخواه است.

[5] Contaminated

قضیه ۱. نقطه فروریزش برآوردگرهای $\hat{\boldsymbol{\beta}}_{FGLTS}(z^*,\lambda)$ و $\hat{\boldsymbol{f}}_{FGLTS}(\boldsymbol{t};z^*,\lambda)$ برابر است با

$$\mathrm{BP}\big(\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*,\lambda)\big)=\frac{n-h-۱}{n}. \qquad (۱۰)$$

برهان. فرض کنید $(\check{\boldsymbol{y}}^*,\check{\boldsymbol{X}}^*)$ نمونه آلوده شده به مشاهدات پرت (نمونه تخریبی) با جایگزینی $m\le n-h$ مشاهده باشد. بنابراین تعداد مشاهدات خوب در $(\check{\boldsymbol{y}}^*,\check{\boldsymbol{X}}^*)$ برابر $n-m\ge h$ بوده و برای نمونه $(\check{\boldsymbol{y}}^*,\check{\boldsymbol{X}}^*)$ داریم

$$\min_{\boldsymbol{z}\in E_h} Q(\boldsymbol{z},۰)=\min_{\boldsymbol{z}\in E_h}\check{\boldsymbol{y}}^{*'}\boldsymbol{S}^{-۱/۲}\boldsymbol{Z}\boldsymbol{S}^{-۱/۲}\check{\boldsymbol{y}}^*\le\min_{\boldsymbol{z}\in E_h}\check{\boldsymbol{y}}'\boldsymbol{S}^{-۱/۲}\boldsymbol{Z}\boldsymbol{S}^{-۱/۲}\check{\boldsymbol{y}}\le hM_{\boldsymbol{y}}^{۲},$$

به‌طوری‌که $M_{\boldsymbol{y}}=\max_{i=۱,\dots,n}|\check{y}_i|$. چنان‌چه $\boldsymbol{\beta}'\boldsymbol{\beta}\ge\frac{hM_{\boldsymbol{y}}^{۲}+۱}{\lambda}$، آن‌گاه (فرض خلف) رابطه

$$\min_{\boldsymbol{z}\in E_h} Q(\boldsymbol{z},\boldsymbol{\beta})\ge\lambda\boldsymbol{\beta}'\boldsymbol{\beta}\ge hM_{\boldsymbol{y}}^{۲}+۱>\min_{\boldsymbol{z}\in E_h} Q(\boldsymbol{z},۰),$$

برقرار است، که چون $\min_{\boldsymbol{z}\in E_h} Q\left(\boldsymbol{z},\hat{\boldsymbol{\beta}}(\boldsymbol{z}^*,k)\right)\le\min_{\boldsymbol{z}\in E_h} Q(\boldsymbol{z},۰)$، فرض خلف باطل می‌شود. بنابراین می‌توان نوشت

$$\hat{\boldsymbol{\beta}}_F(\boldsymbol{z}^*,\lambda)'\hat{\boldsymbol{\beta}}_F(\boldsymbol{z}^*,\lambda)\le\frac{hM_{\boldsymbol{y}}^{۲}+۱}{\lambda},$$

و بنابراین $\mathrm{BP}\left(\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*,\lambda)\right)\ge\frac{n-h-۱}{n}$. اکنون جای آخرین $m=n-h+۱$ مشاهده $(\check{\boldsymbol{X}},\check{\boldsymbol{y}})$ را طوری عوض می‌کنیم که آخرین m مشاهده $(\check{\boldsymbol{X}},\check{\boldsymbol{y}})$ به $((a,۰,\dots,۰)',Ma)$ تغییر یابد، که در آن $M>۰$، $a>۰$ و

$$a^{۲}\ge\max(h-m,۰)(\max_{i=۱,\dots,n}|y_i|+M\max_{i=۱,\dots,n}||\mathbf{x}_i||)^{۲}+\lambda M^{۲}.$$

با فرض $\boldsymbol{\beta}_M=(M,۰,\dots,۰)'\in\mathbb{R}^p$، آخرین m مشاهده $(\check{\boldsymbol{y}}-\check{\boldsymbol{X}}\boldsymbol{\beta}_M)$ برابرند با

$$(Ma-Ma)=۰.$$

بنابراین

$$\min_{\boldsymbol{z}\in E_h} Q(\boldsymbol{z},\boldsymbol{\beta}_M)=\begin{cases}\min\limits_{\boldsymbol{z}\in E_{h-m}}(\check{\boldsymbol{y}}-\check{\boldsymbol{X}}\boldsymbol{\beta}_M)'\boldsymbol{S}^{-۱/۲}\boldsymbol{Z}\boldsymbol{S}^{-۱/۲}(\check{\boldsymbol{y}}-\check{\boldsymbol{X}}\boldsymbol{\beta}_M)+\lambda M^{۲}, & h>M\\ \lambda M^{۲}, & \text{جاهای دیگر}\end{cases}$$

می‌توان نوشت

$$\begin{aligned}\min_{\boldsymbol{z}\in E_h} Q(\boldsymbol{z},\boldsymbol{\beta}_M) &\leq \max(h-m,\circ)(\max_{i=۱,\ldots,n}|\check{y}_i|+M\max_{i=۱,\ldots,n}||\check{\mathbf{x}}_i||)^۲+\lambda M^۲\\ &\leq a^۲-۱. \qquad (۱۱)\end{aligned}$$

همچنین برای نمونه آلوده شده و $\boldsymbol{\beta}=(\boldsymbol{\beta}_۱,\ldots,\boldsymbol{\beta}_p)\in\mathbb{R}^p$، می‌توان نوشت

$$\min_{\boldsymbol{z}\in E_h} Q(\boldsymbol{z},\boldsymbol{\beta})\geq(Ma-a\boldsymbol{\beta}_۱)^۲.$$

زیرا حداقل یکی از m مشاهده آخر در نمونه آلوده متعلق به مجموعه کوچکترین h مانده است. با در نظر گرفتن $\boldsymbol{\beta}'\boldsymbol{\beta}\leq M^۲$، چون $\boldsymbol{\beta}_۱\leq(\boldsymbol{\beta}'\boldsymbol{\beta})^{۱/۲}\leq M$، می‌توان نوشت

$$\min_{\boldsymbol{z}\in E_h} Q(\boldsymbol{z},\boldsymbol{\beta})\geq a^۲(M-\boldsymbol{\beta}_۱)^۲\geq a^۲, \qquad (۱۲)$$

از نابرابری‌های (۱۱) و (۱۲)، نابرابری $\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*,k)'\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*,k)\geq M^۲$ نتیجه می‌شود که به معنی نامتناهی شدن برآوردگر با میل کردن M به بینهایت و فروریزش آن است و بنابراین برهان تمام است.

واضح است که نقطه فروریزش بالاتر به معنی استواری بیشتر برآوردگر در برابر مشاهده‌های دور افتاده است. از رابطه (۱۰) معلوم می‌شود که با کوچک شدن h، نقطه فروریزش بالاتری به‌دست خواهد آمد. به نظر می‌آید که اگر $h<n/۲$ باشد، حتی نقطه فروریزش بیش از ۵۰٪ به دست می‌آید. از لحاظ ریاضی اثبات شده است که نقطه فروریزش هر برآوردگر پایای رگرسیونی حداکثر برابر ۵۰٪ است (روسیو و لروی ۲۰۰۳ را ببینید). با این حال دستیابی عملی به نقطه فروریزش بیش از ۵۰٪ نامعقول است، زیرا اگر بیش از ۵۰٪ مشاهدات از الگوی سایر مشاهدات تبعیت نکنند، اکثریت نقاط در واقع این مجموعه خواهند بود. بنابراین توصیه می‌شود مقادیر $h\geq n/۲$ استفاده نشود. معمولا در عمل مقادیر $\alpha=۰/۷۵$ و یا $h=[[n/۲]]+[[(p+۱)/۲]]$ پیشنهاد می‌شوند.

برای محاسبه $\boldsymbol{Z}^*$ و برآوردگرهای معرفی شده روی مجموعه E_h باید تمامی $\binom{n}{h}$ انتخاب از مجموعه $\{۱,\ldots,n\}$ بررسی شود که به زمان و فضای زیادی نیاز دارد. در این مقاله از روش کم‌ترین توان‌های دوم پیراسته سریع[۶] که توسط روسیو و ون دریسن (۲۰۰۶) معرفی شده استفاده می‌شود.

[6] Fast-LTS

۵ تعیین پارامترهای هموارسازی و اریب‌سازی

برای به‌دست آوردن مقادیر بهینه پارامترهای هموارسازی و مرزبندی شده، بنابر ایوبانک (۱۹۹۹) ابتدا از مقدار بهینه

$$\omega_{\circ}^{*} = n^{-1/5}\Big\{\frac{\kappa_2 \hat{\sigma}_{\circ}^{2}}{M_2^{2} J_2^{2}}\Big\}^{1/5}, \tag{۱۳}$$

برای تعیین پارامتر هموارسازی استفاده می‌شود، که در آن $J_2 = \int_{\circ}^{1} \hat{f}''(t)\,\mathrm{d}t$ اندازه هموار برآورد $f(\cdot)$ بوده و

$$\kappa_2 = \int_{-1}^{1} K(u)^2\,\mathrm{d}u < \infty,\ M_2 = \int_{-1}^{1} u^2 K(u)\,\mathrm{d}u \neq \circ,$$

$$\hat{\sigma}_{\circ}^{2} = \frac{\big\|\boldsymbol{S}^{-1/2}\big(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FG} - \hat{f}_F(\boldsymbol{t})\big)\big\|^2}{n-p},$$

که در آن $\hat{\boldsymbol{\beta}}_{FG} = \big(\check{\boldsymbol{X}}'\boldsymbol{S}^{-1}\check{\boldsymbol{X}}\big)^{-1}\check{\boldsymbol{X}}'\boldsymbol{S}^{-1}\check{\boldsymbol{y}}$ برآورد کم‌ترین توان‌های دوم تعمیم‌یافته محتمل $\boldsymbol{\beta}$ است. اکنون از مقدار بهینه اولیه پارامتر هموارسازی به‌دست آمده در (۱۳) برای تعیین مقدار بهینه اولیه پارامتر مرزبندی شده $(\lambda_{\circ}^{*})$ استفاده می‌شود. این کار براساس معیار اعتبارسنجی تعمیم‌یافته[7]

$$\mathrm{GCV}(\omega_{\circ}^{*}, \lambda) = \frac{\frac{1}{n}\big\|\boldsymbol{S}^{-1/2}\big(\boldsymbol{I}_n - \boldsymbol{L}(\omega_{\circ}^{*}, \lambda)\big)\boldsymbol{y}\big\|^2}{\big[1 - \frac{1}{n}\mathrm{tr}\big(\boldsymbol{L}(\omega_{\circ}^{*}, \lambda)\big)\big]^2},$$

انجام می‌شود، که در آن

$$\boldsymbol{L}(\omega_{\circ}^{*}, \lambda) = \boldsymbol{K} + (\boldsymbol{I}_n - \boldsymbol{K})\boldsymbol{X}(\check{\boldsymbol{X}}'\check{\boldsymbol{X}} + \lambda \boldsymbol{I}_p)^{-1}\check{\boldsymbol{X}}'(\boldsymbol{I}_n - \boldsymbol{K}).$$

ماتریس برازش با استفاده از پهنای باند $\omega_{\circ}^{*}$ است. با این حال یافتن مینیمم تابع GCV با استفاده از رابطه اخیر در هر تکرار از الگوریتم کم‌ترین توان‌های دوم پیراسته سریع مدت زمان اجرای الگوریتم را به گونه قابل توجهی افزایش می‌دهد (واضح است که مینیمم‌سازی تابع GCV خود یک فرایند زمان‌بر است). بنابراین برای برآورد پارامترهای پهنای باند و مرزبندی شده در مراحل بعد، از پارامتر اریب‌سازی پیشنهاد شده توسط هورل و کنارد (۱۹۷۰) استفاده نموده و با l بار تکرار این الگوریتم برآوردگرهای نهایی بهینه استخراج می‌شوند. برای

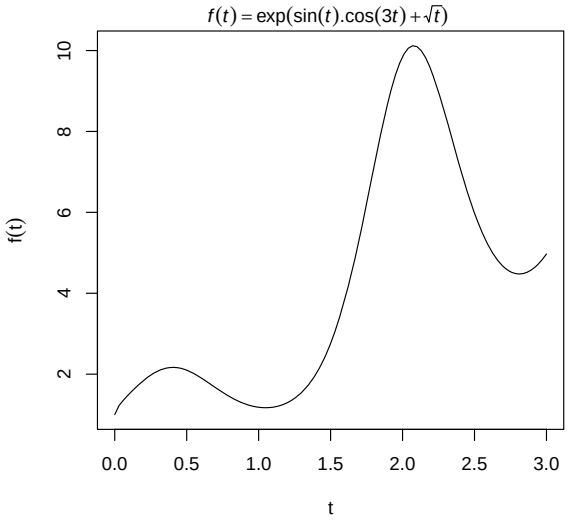
[7] Generalized Cross Validation

$s = ۱, \ldots, l$، در مرحله sام قرار می‌دهیم:

$$\begin{aligned}
\hat{\sigma}_s^۲ &= \frac{\left\| \boldsymbol{S}^{-۱/۲}\big(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*_{s-۱}, \lambda^*_{s-۱}) - \hat{f}_{FGLTS}(\boldsymbol{t}; \boldsymbol{z}^*_{s-۱}, \lambda^*_{s-۱})\big)\right\|^۲}{n - \operatorname{tr}\big(\boldsymbol{L}(\omega^*_{s-۱}, \lambda^*_{s-۱})\big)}, \\
\lambda_s^* &= \frac{p\hat{\sigma}_s^۲}{\left\|\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*_{s-۱}, \lambda^*_{s-۱})\right\|^۲}, \\
\omega_s^* &= n^{-۱/۵}\Big\{\frac{\kappa_۲\hat{\sigma}_s^۲}{M_۲^۲ J_۲^۲}\Big\}^{۱/۵}.
\end{aligned}$$

۶ نتایج عددی

در این بخش برای بررسی کارایی برآوردهای معرفی شده، یک مطالعه شبیه‌سازی به روش مونت کارلو انجام شده و سپس این شیوه برآورد در یک مثال واقعی به کار خواهد رفت. این شبیه‌سازی و تحلیل داده‌های واقعی با استفاده از نرم افزار R نسخه ۲.۳.۳ انجام شده است.



شکل ۱: تابع ناپارامتری در مطالعه شبیه‌سازی

۱.۶ مطالعه شبیه‌سازی به روش مونت کارلو

در این بخش با استفاده از شبیه‌سازی داده‌ها به بررسی و مقایسه برآوردگر پیشنهادی با برآوردگر کلاسیک متداول در مدل نیمه‌پارامتری پرداخته می‌شود. برای رسیدن به ساختار وابستگی، متغیرهای پیشگو بنابر مدل $x_{ij} = (۱ - \gamma^۲)^{\frac{۱}{۲}} z_{ij} + \gamma z_{ip},\ i = ۱, \ldots, n,\ j = ۱, \ldots, p,$ شبیه‌سازی می‌شوند که در آن z_{ij}ها متغیرهای تصادفی نرمال استاندارد بوده و پارامتر γ تعیین کننده میزان همبستگی خطی بین دو متغیر توضیحی است. متغیر پاسخ از مدل رگرسیون نیمه‌پارامتری

$$y_i = \sum_{j=۱}^{۶} x_{ij}\beta_j + f(t_i) + \epsilon_i,\ i = ۱, \ldots, n, \qquad (۱۴)$$

با ۵ متغیر پیشگو برای $n = ۱۵۰$ با $M = ۱۰^۳$ تکرار شبیه‌سازی می‌شود، به طوری که پارامترهای قسمت خطی و غیرخطی مدل (این تابع در شکل ۱ رسم شده است) $\boldsymbol{\beta} = (-۱, ۴, ۲, -۵, -۳)'$ و $f(t) = \exp\left\{\sin(۲t)\cos(۵t) + \sqrt{t}\right\},\ t \in [۰, ۳]$ بوده و خطاها به‌صورت $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_۱^\top, \boldsymbol{\epsilon}_۲^\top)^\top$ هستند، به طوری که

$$\boldsymbol{\epsilon}_{۱\ (h\times ۱)} \sim \mathcal{N}_h(۰, \sigma^۲ \boldsymbol{V}),\ \sigma^۲ = ۱/۴۴,\ v_{ij} = \exp\big(-۹|i-j|\big),$$
$$\boldsymbol{\epsilon}_{۲\ ((n-h)\times ۱)} \overset{i.i.d.}{\sim} \chi^۲_۱(۱۵),$$

که در آن $\chi^۲_\nu(\delta)$ نشان دهنده توزیع کای-دو غیر مرکزی با درجه آزادی ν و پارامتر غیرمرکزی δ است. دلیل اصلی انتخاب این ساختار برای خطاها، ایجاد تعدادی داده پرت به منظور ارزیابی برآوردگر پیشنهادی است. به همین منظور برای $h = [[۰/۶۷n]]$ خطای اول (مشاهدات خوب) از توزیع نرمال و بقیه (مشاهدات پرت) از توزیع مستقل غیرمرکزی کای-دو استفاده شد. غیرمرکزی بودن باعث منحرف شدن داده‌ها به قسمتی دور از مدل رگرسیون شده که در نتیجه برآورد غیراستوار را به سمت خود منحرف می‌کنند.

نتایج در جدول ۱ برای $n = ۱۵۰$ و $\gamma = ۰/۹۰, ۰/۹۵$ با ۳۳ درصد نقاط دورافتاده گزارش شده است. در این جدول مقدار برآورد میانگین توان‌های دوم خطای برآوردگرها به‌صورت

$$\widehat{\text{mse}}\big(\hat{\boldsymbol{\beta}}\big) = \frac{۱}{M}\sum_{j=۱}^{M}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}),$$
$$\widehat{\text{mse}}\big(\hat{f}(\boldsymbol{t})\big) = \frac{۱}{M}\sum_{j=۱}^{M}(\hat{f}_j(\boldsymbol{t}) - f(\boldsymbol{t}))^۲,$$

جدول ۱: مقایسه برآوردگرهای پیشنهادی به روش مونت کارلو

برآوردگر	$\gamma = ۰/۹۰$		$\gamma = ۰/۹۵$	
	$\hat{\boldsymbol{\beta}}_{FG}$	$\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*, \lambda^*)$	$\hat{\boldsymbol{\beta}}_{FG}$	$\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*, \lambda^*)$
	−۱/۰۰۶۴	−۱/۰۰۱۲	−۱/۰۱۱۹	−۱/۰۰۲۵
	۳/۸۵۸۸	۳/۹۷۴۱	۳/۷۳۷۷	۳/۹۴۴۰
میانگین ضرایب	۱/۵۷۶۴	۱/۹۲۲۳	۱/۲۱۳۱	۱/۸۳۲۰
	−۴/۶۳۴۲	−۴/۹۳۲۹	−۴/۳۲۰۴	−۴/۸۵۴۹
	−۲/۸۰۷۵	−۲/۹۶۴۷	−۲/۶۴۲۳	−۲/۹۲۳۶
$\widehat{\mathrm{mse}}(\hat{\boldsymbol{\beta}})$	۱/۱۷۷۶	۰/۳۶۷۱	۲/۳۵۷۱	۰/۹۹۰۳
$\widehat{\mathrm{mse}}(\hat{f}(\boldsymbol{t}))$	۰/۵۵۶۲	۰/۲۹۵۹	۰/۴۷۲۹	۰/۳۱۳۹

به‌دست می‌آید، که در آن $M = ۱۰^۳$ تعداد تکرارهای شبیه‌سازی است. همان‌طور که در جدول ۱ دیده می‌شود، برآوردگرهای قسمت خطی و غیرخطی استوار مرزبندی شده دارای میانگین توان دوم خطای کمتری نسبت به برآوردگرهای غیراستوار هستند. همچنین کمتر بودن اریبی برآوردگرهای قسمت خطی استوار مرزبندی شده به خوبی در این جدول قابل‌ملاحظه است.

۲.۶ تحلیل داده‌های واقعی

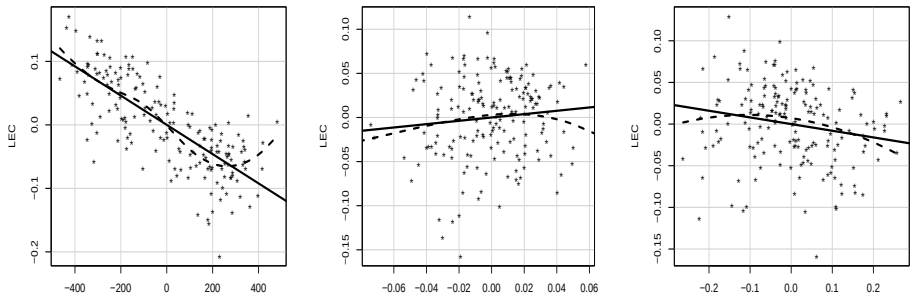
مجموعه داده‌ها مربوط به مصرف برق است که توسط آکدنیز دوران و همکاران(۲۰۱۲) مطالعه شده است. متغیرها برای $n = ۱۷۷$ مورد به این ترتیب تعریف می‌شوند: متغیر پاسخ LEC لگاریتم مصرف ماهیانه برق به ازای هر نفر و متغیرهای پیشگو شامل LI لگاریتم درآمد هر فرد، $LREG$ لگاریتم نسبت قیمت برق به قیمت گاز و $Temp$ جمع شاخص متوسط دما (برای یک ماه خاص) هستند. داده‌ها مربوط به ۲۰ شهر از شهرهای آلمان است که از سازمان هواشناسی آلمان به‌دست آمده‌اند. در اینجا با استفاده از نمودار متغیر افزوده[8] بخش ناپارامتری مدل را تشخیص دهیم. نمودار متغیر افزوده به طور شهودی اثر هر یک از متغیرهای پیشگو را پس از حذف اثر سایر متغیرهای پیشگو، بر متغیر پاسخ آشکار می‌کند. با نگاه کردن به نمودار متغیر افزوده در

[8] Added Variable Plot

شکل ۲ از آن‌جا که شواهدی مبنی بر این‌که رابطه جزئی میان متغیر پاسخ مصرف برق و متغیر جمع شاخص میانگین دما $(Temp)$ یک رابطه غیرخطی است از نمودار به دست می‌آید، این متغیر بعنوان جزء ناپارامتری مدل در نظر گرفته می‌شود، بنابراین مدل رگرسیون نیمه‌پارامتری به‌صورت

$$(LEC)_i = \beta_1(LI)_i + \beta_2(LREG)_i + f(Tempe)_i + \epsilon_i, \qquad (۱۵)$$

است، که در آن ضریب همبستگی بین متغیرهای پیش‌بین قسمت خطی برابر $-۰/۵۶۹۷$ و نشان‌دهنده وجود هم‌خطی بین آن‌ها است. همچنین شکل ۲ وجود مشاهده‌های دورافتاده زیادی را نشان می‌دهد که می‌توانند بر برآوردهای مرزبندی شده غیر استوار اثر بگذارند. شماره مشاهده‌های دورافتاده که تاثیر مخرب جدی بر برآوردگر کم‌ترین توان‌های دوم می‌گذارند عبارتند از: ۳، ۱۰، ۱۵، ۱۸، ۲۰، ۲۷، ۶۱، ۱۱۱، ۱۲۱، ۱۲۳، ۱۳۳، ۱۳۹، ۱۶۰ و ۱۷۲ که حدود ۸ درصد از کل مشاهدات را تشکیل می‌دهند.



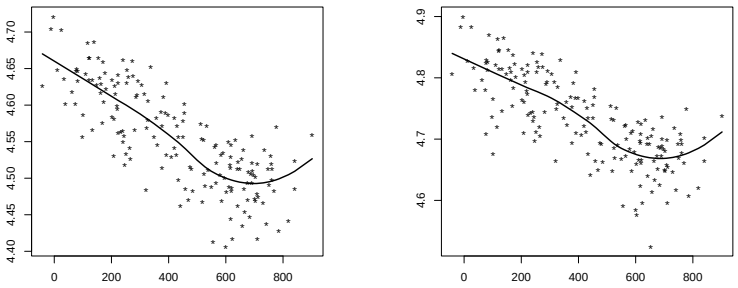
(ج) (ب) (الف)

شکل ۲: نمودار افزوده متغیرهای پیشگو در مقابل متغیر پاسخ و برازش کم‌ترین توان‌های دوم (خط ممتد) و برازش ناپارامتری کرنل (خط چین)، الف- لگاریتم درآمد هر فرد، ب- لگاریتم نسبت قیمت برق به گاز، ج- میانگین دما

در جدول ۲ مقادیر برآوردگرها و همچنین RSS و R^2 که به ترتیب، مجموع توان‌های دوم مانده‌ها و ضریب تغییرات مدل رگرسیون نیمه‌پارامتری برازش شده هستند، گزارش شده‌اند. همان‌طورکه در این جدول دیده می‌شود، برآوردگرهای استوار مرزبندی شده محتمل دارای مجموع توان دوم خطای کمتری نسبت به برآوردگرهای غیر استوار محتمل بوده و حدود ۴ درصد بیشتر تغییرات متغیر پاسخ را نسبت به برآوردگر غیر استوار بیان می‌کنند. تابع ناپارامتری برازش شده در شکل ۳ رسم شده است.

جدول ۲: مقایسه برآوردگرهای پیشنهادی در داده‌های واقعی

برآوردگر	$\hat{\boldsymbol{\beta}}_{FG}$	$\hat{\boldsymbol{\beta}}_{FGLTS}(\boldsymbol{z}^*, \lambda^*)$
$\hat{\beta}_۱$	۰/۱۴۳۰	۰/۱۸۸۷
$\hat{\beta}_۲$	−۰/۰۸۱۱	−۰/۰۷۹۵
RSS	۰/۳۴۳۳	۰/۲۷۲۱
R^2	۰/۶۳۱۹	۰/۶۷۴۱



(الف) (ب)

شکل ۳: الف: برازش تابع ناپارامتری به روش غیراستوار ب: برازش تابع ناپارامتری به روش استوار مرزبندی‌شده

بحث و نتیجه‌گیری

در این مقاله برآورد مرزبندی شده استوار محتمل ضرایب خطی و توابع غیرخطی در مدل رگرسیون نیمه‌پارامتری درحضور داده‌های پرت و هم‌خطی چندگانه به کمک الگوریتم سریع کم‌ترین توان‌های دوم مرزبندی شده پیراسته با فرض همبسته بودن جملات خطا، ارائه شد. نقطه فروریزش برآوردگرهای معرفی شده به دست آمد و از لحاظ نظری نشان داده شد که این نقطه فروریزش می‌تواند بیش از ۵۰ درصد باشد. با این حال از دیدگاه عملی دستیابی به نقطه فروریزش بیش از ۵۰ درصد امکان‌پذیر نیست.

نتایج عددی مطالعه شبیه‌سازی نشان می‌دهد که استفاده از برآوردگر استوار مرزبندی شده تعمیم‌یافته محتمل در مدل رگرسیون نیمه‌پارامتری، دارای تاثیر به سزایی در بهبود برآوردگرهای ضرایب خطی و توابع غیرخطی مدل در برازش مدل به داده‌های همبسته و دارای مشاهدات دورافتاده و دارای هم‌خطی چندگانه است.

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

نوروزی‌راد، م. و آرشی، م. (۱۳۹۶)، مطالعه رفتار حدی برآوردگرهای انقباضی در مدل رگرسیون تاوانیده با نرم مستطیلی، مجله علوم آماری ایران، **۱۱**، ۱۴۹-۱۷۴.

Akdeniz Duran, E., Hardle, W. K. and Osipenko, M. (2012), Difference-based Ridge and Liu Type Estimators in Semiparametric Regression Models, *Journal of Multivariate Analysis*, **105**, 164-175.

Arashi, M., Janfada, M. and Norouzirad, M. (2015), Singular Ridge Regression with Stochastic Constraints, *Communication in Statistics - Theory and Methods*, **44**, 1281-1292.

Arashi, M. and Norouzirad, M. (2015), Improved Ridge M-estimators, 46th Annual Iraninan Mathematics Conference, Yazd University, Yazd, Iran.

Arashi, M. and Valizadeh, T. (2015), Performance of Kibria's Methods in Partial Linear Ridge Regression Model, *Statistical Papers*, **56**, 231-246.

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**, 226-248.

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

Blanchfower, D. G. and Oswald A. J. (1994), *The Wage Curve*, MIT Press Cambridge.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986), Semiparametric Estimates of the Relation Between Weather and Electricity Sales, *Journal of the American Statistical Association*, **81**, 310-320.

Eubank, R.L. (1999), *Nonparametric Regression and Spline Smoothing*, Second edition, Marcel Dekker, New York.

Farebrother, R. (1976), Further Results on the Mean Square Error of Ridge Regression, *Journal of the Royal Statistical Society* Ser. B, **38**, 248-250.

Golub, G., Heath, M. and Wahba, G. (1979), Generalized Cross Validationas a Method for Choosing a Good Ridge Parameter, *Technometrics*, **21**, 215-223.

Hardle, W., Liang, H. and Gao, J. (2000), *Partially Linear Models*. Physika Verlag, Heidelberg.

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

Gao, J. T., Hong, S. Y. and Liang, H. (1995), Convergence Rates of a Class of Estimates in Partly Linear Models, *Acta Mathematica Sinica*, **38**, 658-669.

Maronna, R. A., Martin, D. R. and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, 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 Babaie-Kafaki, S. (2016), Extended Least Trimmed Squares Estimator in Semiparametric Regression Models with Correlated Errors, *Journal of Statistical Computation and Simulation*, **186**, 357-372.

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. (2003), *Robust Regression and Outlier Detection*, 2nd ed. Wiley, Hoboken.

Rousseeuw, P. J. and van Driessen, K. (2006), Computing LTS Regression for Large Data Sets, *Data Mining and Knowledge Discovery* **12**, 29-45.

Speckman, P. (1988), Kernel Somoothing in Partial Linear Models, *Journal of the Royal Statistical Society* Ser. B., **50**, 413-436.

Tabakan, G. and Akdeniz, F. (2010), Difference-based Ridge Estimator of Parameters in Partial Linear Model, *Statistical Papers*, **51**, 357-368.

Turkmen, A. S. and Tabakan, G. (2015), Outlier Resistant Estimation in Difference-based Semiparametric Partially Linear Models, *Communications in Statistics - Simulation and Computation*, **44**, 417-432.

Willis, R. J. (1986), Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions in: Ashenfelter, O. and Layard, R. The Handbook of Labor Economics, Vol.1 *North Holland-Elsevier Science Publishers Amsterdam*, **1**, 525-602.

Zellner, A. (1962), An Effivient Method of Estimating Seemingly Unrelated Regression and Tests for Aggregation Bias, *Journal of the American Statistical Association*, **58**, 977-992.

Zellner, A. (1963), Estimators for Seemingly Unrelated Regressions: Some Finite Sample Results, *Journal of the American Statistical Association*, **58**, 977-992.

Feasible Generalized Rdge Robust Estimator in Semiparametric Regression Models

Mahdi Roozbeh[1], **Morteza Amini**[2]

[1]Department of Statistics, Semnan University, Semnan, Iran.

[2]Department of Statistics, University of Tehran, Tehran Iran.

Abstract: In many fields such as econometrics, psychology, social sciences, medical sciences, engineering, etc., we face with multicollinearity among the explanatory variables and the existence of outliers in data. In such situations, the ordinary least-squares estimator leads to an inaccurate estimate. The robust methods are used to handle the outliers. Also, to overcome multicollinearity ridge estimators are suggested. On the other hand, when the error terms are heteroscedastic or correlated, the generalized least squares method is used. In this paper, a fast algorithm for computation of the feasible generalized least trimmed squares ridge estimator in a semiparametric regression model is proposed and then, the performance of the proposed estimators is examined through a Monte Carlo simulation study and a real data set.

Keywords: Breakdown point, Generalized cross validation, Least trimmed squares estimator, Outliers, Semiparametric regression model.

Mathematics Subject Classification (2010): 62G08، 62G35، 62J20.