مدل‌های نیمه‌پارامتری استوار تنک در داده‌های با بعد بالا

مهدی روزبه[۱]، منیره معنوی[۲]

تاریخ دریافت: ۱۴۰۰/۰۴/۲۷
تاریخ پذیرش: ۱۴۰۰/۱۲/۲۸

چکیده:
تحلیل و مدل‌بندی داده‌های با بعد بالا یکی از چالش‌برانگیزترین مسائل روز دنیا است. تفسیر این داده‌ها کاری ساده نیست و نیازمند استفاده از روش‌های مدرن است. روش‌های تاوانیده یکی از رایج‌ترین راه‌های تحلیل داده‌های با بعد بالاست. همچنین مدل‌های رگرسیونی و تحلیل آن‌ها به‌شدت تحت تأثیر مشاهدات پرت قرار می‌گیرند. روش کمترین توان‌های دوم پیراسته یکی از بهترین روش‌های استوار برای از بین بردن تأثیر تخریبی این نقاط است. مدل‌های نیمه‌پارامتری که ترکیبی از هر دو نوع مدل‌های پارامتری و ناپارامتری هستند، بسیار انعطاف‌پذیرند. این مدل‌ها زمانی مفید هستند که هم بخش پارامتری و هم بخش ناپارامتری در مدل وجود دارد. هدف اصلی این مقاله تحلیل مدل‌های نیمه پارامتری در داده‌های با بعد بالا با حضور نقاط پرت با استفاده از روش لاسو تنک استوار است. در انتها، کارایی برآوردگر پیشنهادی با استفاده از تحلیل داده‌هایی واقعی در مورد تولید ویتامین B2 سنجیده می‌شود.

واژه‌های کلیدی: داده‌های با بعد بالا، روش لاسو، روش کمترین توان‌های دوم پیراسته، روش کمترین توان‌های دوم پیراسته تنک، مدل‌های نیمه‌پارامتری.

۱ مقدمه

امروزه با گسترش روزافزون علم، دانش و فناوری، روش‌های نوین و دقیقی برای اندازه‌گیری، جمع‌آوری و ثبت اطلاعات ابداع شده و این امر باعث ظهور و گسترش داده‌های بعد بالا شده است. از اواخر سال ۱۹۹۰ کار با این مجموعه داده‌ها یعنی داده‌هایی که در آن تعداد متغیرهای توضیحی (p) بسیار بیشتر از تعداد مشاهدات (n) است، شروع شد [۱۱]. افزایش بیش‌ازپیش ابعاد داده‌ها به یک مسئله اساسی در مبحث داده‌کاوی تبدیل‌شده است. تجزیه، تحلیل و تفسیر این داده‌ها امری دشوار و بسیار قابل‌تأمل است که واسرمن [۲۶] در کتاب خود به این امر اشاره نموده و از این پدیده تحت عنوان نفرین ابعاد[۳] یاد کرده است. اصطلاحی که معمولاً به بلمن نسبت داده می‌شود. تقریباً این بدان معنی است که با افزایش بعد مشاهدات، برآورد پارامترها با سرعت زیادی، بسیار دشوار می‌شود. حداقل دو نوع از این نفرین وجود دارد. اولین مورد، نفرین محاسباتی ابعاد است. این امر به این واقعیت اشاره دارد که بار محاسباتی برخی از روش‌ها می‌تواند به‌صورت نمایی با بعد افزایش یابد. دومین مورد که واسرمن آن را نفرین آماری ابعاد نامید به این صورت است که اگر داده‌ها دارای بعد d باشند، به نمونه‌ای با اندازه n موردنیاز است که به‌صورت نمایی با d رشد کند.

درگذشته، کار با داده‌های کلاسیک که دارای حداکثر چند ده متغیر توضیحی بوده‌اند، بسیار ساده بود و روش‌های کلاسیک نظیر کمترین توان‌های دوم نتایج قابل قبولی ارائه می‌دادند؛ اما این روش‌ها در حضور داده‌های بعد بالا برآوردی ارائه نمی‌دهند. چراکه برای محاسبه برآوردگر کمترین توان‌های دوم یعنی $\hat{\beta}_{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ نیاز به محاسبه وارون ماتریس $\boldsymbol{X}^\top \boldsymbol{X}$ است. ولی این ماتریس در این وضعیت رتبه کامل نبوده و درنتیجه وارون‌پذیر نخواهد بود.

بنابراین استفاده از روش‌های کلاسیک در این شرایط سودمند نخواهد بود و مستلزم به‌کارگیری روش‌های دیگری خواهیم بود. در تحلیل رگرسیونی داده‌های بعد بالا تعداد زیاد متغیرهای توضیحی، محقق را با چالشی جدی روبه‌رو می‌کند و همیشه جدالی بین دقت،

[۱] هیئت‌علمی گروه آمار، دانشگاه سمنان، سمنان، ایران (نویسنده مسئول:mahdi.roozbeh@semnan.ac.ir)
[۲]دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.

[3]Curse of dimensionality
[4]Overfiting

سرعت و هزینه وجود دارد. از سوی دیگر، وجود متغیرهایی در مدل که با متغیر پاسخ ارتباطی ندارند، منجر به بیش‌برازش[۴] می‌شود. چنین مدلی، مدلی بسیار پیچیده برای داده‌ها است. در این شرایط، مدل با تغییرات جهشی سعی در پوشش داده‌های حاصل از نمونه و حتی مقدارهای خطاها می‌کند. درحالی‌که این مدل باید منعکس‌کننده‌ی رفتار جامعه باشد. به‌کارگیری تمامی متغیرهای توضیحی زمانی طولانی صرف کرده و هزینه‌های محاسباتی بسیاری را بر محقق تحمیل می‌کند. در این شرایط این ابهام مطرح می‌شود که شاید حضور تمامی متغیرها برای برازش مدل الزامی نباشد و بدین‌سان اغلب محققان با در نظر گرفتن فرض تنکی[5] سعی در کاهش بعد دارند و از زیرمجموعه‌ای مناسب از متغیرها برای برازش مدل کمک می‌گیرند و ازاین‌رو علاقه‌مند به شناسایی این زیرمجموعه مناسب هستند. روش‌های متنوعی برای رویارویی با داده‌های با بعد بالا وجود دارد. روش کمترین توان‌های دوم تاوانیده[6] یکی از روش‌های بسیار مفید برای تحلیل داده‌های با بعد بالا است.

نقاط پرت یکی دیگر از معضلات مدل‌های رگرسیونی است. بارنت و لوئیس [۹] نقطه پرت را مشاهده‌ای (مجموعه‌ای از مشاهدات) تعریف کردند که به نظر می‌رسد دارای باقی‌مانده‌ای متفاوت با باقی‌مانده‌ی سایر نقاط داده‌ها باشد. این نقاط، تأثیرات بسیار بدی روی عملکرد مدل‌های رگرسیونی دارند. روش‌های استوار[7]، روش‌هایی هستند که برای از بین بردن این اثرات استفاده می‌شوند. روش کمترین توان‌های دوم پیراسته[8] یکی از مشهورترین روش‌های استوار است که در این مقاله از آن استفاده می‌شود. این روش توسط روسو و لروی [۲۲] معرفی شد.

مدل‌های پارامتری با تمام مزایا و محاسن خود، نمی‌توانند نماینده خوبی از آنچه در دنیای واقعی رخ می‌دهد، باشند. در دنیای واقعی به‌ندرت می‌توان مجموعه داده‌هایی که در آن ارتباط خطی بین متغیر توضیحی و متغیر پاسخ وجود داشته باشد، یافت. در عمل مجموعه داده‌های دنیای واقعی بسیار پیچیده و دارای ارتباطات غیرخطی‌اند که در مواردی حتی با به‌کارگیری تبدیل‌های قوی نظیر تبدیل باکس-کاکس، لگاریتمی و ... نیز ارتباط مابین آن‌ها خطی نمی‌شود. در این شرایط ناگزیر به استفاده از روش‌های دیگری خواهیم بود.

مدل‌های رگرسیونی را می‌توان به سه دسته پارامتری، نیمه‌پارامتری[9] و ناپارامتری[10] تقسیم کرد. مدل‌های رگرسیونی پارامتری صرفاً به‌منظور بررسی ارتباطات خطی میان متغیر پاسخ و متغیرهای توضیحی به کار گرفته می‌شوند. بدین سبب کار کردن با آن‌ها بسیار ساده و لذت‌بخش است و تحلیل نتایج آن پیچیدگی عجیبی ندارد. مدل‌های رگرسیونی ناپارامتری زمانی که تمام متغیرهای توضیحی ارتباط غیرخطی با متغیر پاسخ دارند، مورداستفاده قرار می‌گیرند. مدل‌های ناپارامتری ممکن است، نماینده بهتری برای مجموعه داده‌های واقعی باشند، اما متأسفانه تجزیه‌وتحلیل آن‌ها کار ساده‌ای نیست. مدل‌های نیمه‌پارامتری از دو قسمت پارامتری و ناپارامتری تشکیل‌شده‌اند. مدل‌های رگرسیونی نیمه‌پارامتری انعطاف‌پذیرتر از مدل‌های رگرسیونی ناپارامتری و پارامتری هستند زیرا هم ویژگی‌های مدل رگرسیونی پارامتری را دارا هستند و هم ویژگی‌های مدل رگرسیونی ناپارامتری [۳]. به‌عبارت‌دیگر مدل رگرسیونی نیمه‌پارامتری انعطاف‌پذیری یک مدل رگرسیونی ناپارامتری و قدرت توضیحی یک مدل رگرسیونی پارامتری را حفظ می‌کند. انگل و همکارانش [۱۲] برای نخستین بار این مدل‌ها را به‌منظور مدل‌سازی مصرف ماهیانه برق خانوارها ابداع کردند و به‌طور رسمی آن را در آمار به کار بردند. متغیرهای موردبررسی آن‌ها مصرف برق ماهیانه خانوارها، درآمد ماهیانه خانوارها، قیمت ماهیانه برق و دمای هوا بود. متغیر مصرف ماهیانه برق خانوارها متغیر پاسخ و سایر متغیرها، توضیحی هستند. در بررسی‌های اولیه انگل و همکارانش دریافتند که رابطه بین درجه حرارت و مصرف ماهیانه برق بسیار غیرخطی است؛ زیرا مصرف ماهیانه برق هم در درجه حرارت پایین و هم در درجه حرارت بالا افزایش می‌یابد که این ارتباط قابل‌چشم‌پوشی نبوده چراکه این ارتباط بسیار مهم بوده و فروش ماهیانه برق در شرایط عادی هوا بسیار متفاوت از شرایط غیرعادی بوده است. از سوی دیگر ارتباط میان مصرف ماهیانه برق و درآمد خانوارها و همچنین ارتباط میان مصرف ماهیانه برق و قیمت ماهیانه برق نیز خطی بوده است. به همین سبب برای مدل‌سازی این مورد نه استفاده از مدل‌های پارامتری و نه به‌کارگیری مدل‌های ناپارامتری راهگشا نبود. انگل و همکارانش در این شرایط وادار به استفاده از روش دیگری بودند و این امر باعث ظهور مدل‌های نیمه‌پارامتری شد. این مدل‌ها به‌سرعت جایگاه خود را در حیطه‌های مختلف علوم یافتند به‌طوری‌که کاربردهایشان در علوم اقتصادی و تحلیل بقا زبانزد است. مدل‌های نیمه‌پارامتری پیشرفت

[5]Sparsity
[6]Penalized least squares method
[7]Robust methods
[8]Least trimmed squares method
[9]Semiparametric
[10]Nonparametric

بسیار سریعی داشته و موردتوجه بسیاری از محققان و کارشناسان آمار قرار گرفته است. اسپکمن [۲۳] برآورد باقی‌مانده‌ی جزئی پارامتر قسمت خطی و تابع ناپارامتری، واریانس و اریبی مجانبی برآوردگرها را به دست آورد. بونیا [۱۰] روش‌های انتخاب متغیرهای کمکی سازگار را در مدل‌های نیمه‌پارامتری بر اساس روش کمترین توان‌های دوم تاوانیده پیشنهاد نمود. او نشان داد که برآوردگر انتخاب‌شده از قسمت خطی به‌طور مجانبی نرمال است. آکدنیز و همکاران [۴] نیز با استفاده از این مدل‌ها اقدام به مدل‌سازی مصرف ماهیانه برق در حضور متغیرهای توضیحی درآمد، دمای هوا و نسبت نرخ برق به گاز کردند.
روزبه و آرشی [۲۱]، آرشی و همکاران [۷]، امینی و روزبه [۶]، روزبه [۱۹، ۲۰]، یوزباشی و همکاران [۲۵] و روزبه و چاچی [۳] در این حیطه تحقیقات ارزشمندی انجام دادند.
برای مشاهده کامل مدل‌های نیمه‌پارامتری و کاربردهایش کتاب هاردل [۱۶] منبع بسیار خوبی است که به خواننده پیشنهاد می‌شود.

در بخش ۳ روش کمترین توان‌های دوم تاوانیده معرفی می‌شود. روش کمترین توان‌های دوم پیراسته در بخش ۴ بیان می‌شود. در بخش ۵ مدل نیمه‌پارامتری کمترین توان‌های دوم پیراسته روی داده‌های مربوط به ریبوفلاوین به کار گرفته می‌شود.

۲ مدل‌های نیمه پارامتری

مدل رگرسیونی نیمه‌پارامتری به‌صورت

$$\boldsymbol{y} = \boldsymbol{X\beta} + f(\mathbf{t}) + \varepsilon, \qquad (1)$$

تعریف می‌شود که در آن

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, f(\mathbf{t}) = \begin{bmatrix} f(t_1) \\ f(t_2) \\ \vdots \\ f(t_n) \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

که متغیر پاسخ $\boldsymbol{y}$، ماتریس طرح $\boldsymbol{X}$، بردار پارامتر مجهول $\boldsymbol{\beta}$، $f(.)$ تابع حقیقی مقدار مجهول از مقادیر متغیر $\mathbf{t}$، و بردار خطا ε با شرایط

$$E(\boldsymbol{\varepsilon}) = 0, \quad E(\boldsymbol{\varepsilon\varepsilon}^\top) = \sigma^2 \mathbf{I}_{(p+1)\times(p+1)}$$

است. در مدل (۱) متغیر پاسخ $\boldsymbol{y}$ به‌طور خطی با متغیرهای توضیحی $\mathrm{x}_1, \ldots, \mathrm{x}_p$ در ارتباط است و با متغیر توضیحی $\boldsymbol{t}$ ارتباط غیرخطی دارد. واضح است که به دلیل دو قسمتی بودن این مدل‌ها (پارامتری و ناپارامتری) مسئله برآورد پارامترها کمی پیچیده خواهد شد. در میان کتب و مقالات آماری روش‌های متفاوتی به‌منظور برآورد پارامترهای این مدل‌ها توسط محققان مختلف ارائه‌شده است. یکی از روش‌های جالب و پرکاربرد روش پیشنهادی فن و همکارانش [۱۴] است. او و همکارانش طی دو مرحله پارامترهای مدل را برآورد نمودند. به‌این‌ترتیب که در مرحله اول قسمت پارامتری مدل و در مرحله دوم قسمت ناپارامتری مدل برآورد می‌شود. قدم اول به‌منظور برآورد پارامترها به روش فن [۱۴] معلوم در نظر گرفتن بردار پارامتر $\boldsymbol{\beta}$ است. با این فرض می‌توان نوشت

$$\boldsymbol{y} - \boldsymbol{X\beta} = f(\boldsymbol{t}) + \varepsilon \qquad (2)$$

اکنون با بازنویسی رابطه (۲) خواهیم داشت

$$\mathring{\boldsymbol{y}} = f(\boldsymbol{t}) + \boldsymbol{\epsilon} \qquad (3)$$

به‌طوری‌که در آن $\mathring{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X\beta}$ است. بدیهی است که مدل (۳) یک مدل ناپارامتری است. کافی است که تابع f برآورد شود؛ اما برآورد قسمت ناپارامتری خود امری بسیار مهم و حائز اهمیت است. روش‌های گوناگونی به‌منظور ارائه قسمت ناپارامتری و تجزیه‌وتحلیل مدل‌های ناپارامتری وجود دارد. در این مقاله از روش کرنل برای برآورد قسمت ناپارامتری استفاده شده است. تابع کرنل یک تابع تک مدی است که بیشترین وزن را به نقطه موردنظر (نقطه‌ای که قرار است در آن تابع برآورد شود)، در اینجا t_0 می‌دهد. غالباً تابع کرنل یک تابع متقارن است که به فاصله‌های مساوی از هر دو سمت وزن یکسان اختصاص می‌دهد، البته این امر یک قاعده کلی نیست و $K(.)$ می‌تواند وزن‌های غیر یکسان نیز توزیع نماید. کرنل یک برآوردگر اریب با واریانس متناهی است. کرنل $K(.) \geq 0$ در حقیقت تابع همواری است که سه ویژگی

$$\int K(t)d(t) = 1, \int tK(t)d(t) = 0, \; 0 < \sigma_t^2 \equiv \int t^2 K(t)dt < \infty$$

را دارا باشد.

با ظهور کرنل، کاربرد فراوان و محبوبیت آن سبب پیدایش توابع مختلف شد. از محبوب‌ترین کرنل‌ها می‌توان به کرنل یکنواخت

(مستطیلی)[11]، کرنل مثلثی[12]، کرنل درجه چهارم[13]، کرنل اپاشینکو[14]، کرنل سه وزنی[15]، کرنل سه مکعبی[16]، کرنل کسینوسی[17]، کرنل لاپلاسین گوسی[18] و کرنل نرمال (گاوسی)[19] اشاره کرد.

همان‌طورکه ذکر شد هدف یافتن برآورد $f(.)$ است. با فرض هموار بودن (دارا بودن مشتق اول و دوم) $f(.)$ می‌توان نوشت

$$\hat{f}(t) = \sum_{i=1}^{n} k_i(t)(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$$

که در آن $k_i(.)$ تابع کرنل و $\mathbf{x}_i^\top$ سطر i ام ماتریس $\mathbf{X}$ است. در حقیقت در این روش برای برآورد قسمت ناپارامتری از میانگین وزنی استفاده می‌شود. اکنون به‌منظور برآورد بردار ضرایب رگرسیونی $\boldsymbol{\beta}$، $\hat{f}$ در مدل (۱) جایگذاری می‌شود بدین ترتیب به‌سادگی می‌توان بردار پارامترها را برآورد نمود. به‌عبارت‌دیگر می‌بایست مدل را (۱) به‌صورت

$$\boldsymbol{y} - \hat{f}(\boldsymbol{t}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{۴}$$

بازنویسی نمود. اکنون مدل (۴) یک مدل پارامتری است که در درون خود اثر ناپارامتری را پنهان نموده است. با ساده‌سازی مدل (۴) خواهیم داشت:

$$y_i - \sum_{j=1}^{n} k_j(t_i) y_j = (\mathbf{x}_i^\top - \sum_{j=1}^{n} k_j(t_i)\mathbf{x}_j^\top)\boldsymbol{\beta} + \varepsilon_i$$
$$\Rightarrow \tilde{y}_i = \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n.$$

که در آن $\tilde{y}_i = y_i - \sum_{j=1}^{n} k_j(t_i) y_j$ و $\tilde{\mathbf{x}}_i^\top = \mathbf{x}_i^\top - \sum_{j=1}^{n} k_j(t_i)\mathbf{x}_j^\top$ است. با در نظرگرفتن فرم ماتریسی مدل به‌صورت $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ برآورد ضرایب رگرسیونی با استفاده از روش کمترین توان‌های دوم معمولی به‌صورت زیر است.

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}.$$

۳ روش کمترین توان‌های دوم تاوانیده

با ظهور داده‌های با بعد بالا و ناتوانی روش‌های کلاسیک برای تحلیل و تفسیر مدل‌های رگرسیونی نیاز به معرفی روش‌های جدیدی احساس می‌شد. همان‌طور که در بخش قبل مطرح شد روش محبوب کمترین توان‌های دوم در مواجهه با داده‌های با بعد بالا قادر به محاسبه برآورد ضرایب نبود و بدین‌سان ایده تاوانیدن این برآوردگر میان محققان آماری شکل گرفت. در واقع برآوردگرهای تاوانیده حاصل از به‌کارگیری روش‌های کمترین توان‌های دوم یا ماکسیمم درستنمایی تاوانیده در مدل‌بندی رگرسیونی است. در حقیقت این برآوردگرها از روش بهینه‌سازی یک تابع درجه دوم (البته نه همیشه) نسبت به یک برآوردگر تاوانیده به دست می‌آید. ایده اصلی در این برآوردگرها، این است که ضرایب برآورد شده متغیرهای توضیحی کم‌اهمیت در یک مدل تنک، به سمت صفر منقبض شود. در اکثر روش‌های تاوانیده، این متغیرها از مدل خارج می‌شوند و همین امر موجب ساده‌تر شدن تفسیرپذیری مدل خواهد شد [۱].

مسئله کمینه‌سازی در روش کمترین توان‌های دوم معمولی در مدل فوق به‌صورت

$$\min_{\boldsymbol{\beta}} \Big\{ (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})^\top (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) \Big\}, \tag{۵}$$

است. با محاسبه مشتق تابع هدف عبارت (۵) نسبت به بردار پارامتر $\boldsymbol{\beta}$ و برابر صفر قرار دادن آن و حل معادله حاصل، برآوردگر کمترین توان‌های دوم به‌سادگی یافت می‌شود. به‌عبارت‌دیگر در این روش خطا برحسب $\boldsymbol{\beta}$ کمینه می‌شود. واضح است که اگر مقادیر واقعی $\boldsymbol{\beta}$ بزرگ باشد، برآوردگر حاصل، فارغ از احتساب بزرگی $\boldsymbol{\beta}$، فاصله زیادی از مقدار واقعی $\boldsymbol{\beta}$ خواهد داشت و به همین سبب خطای برآورد زیاد شده و دقت آن کاهش می‌یابد [۱]. یک روش بسیار سودمند و تقریباً ساده برای مقابله با این مشکل، تاوانیدن مقادیر بزرگ $\boldsymbol{\beta}$ است بدین‌سان به‌جای کمینه کردن تابع هدف در روش کمترین توان‌های دوم، مسئله کمینه‌سازی تاوانیده یعنی

$$\min_{\boldsymbol{\beta}} \Big\{ (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})^\top (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) \Big\}, \quad s.t. \quad p(\boldsymbol{\beta}) \leq t, \tag{۶}$$

جایگزین می‌شود. عدد ثابت $t \geq 0$ پارامتر تنظیم‌کننده[20] نامیده می‌شود که اگر برابر صفر باشد، مدل تنها شامل عرض از مبدأ است و اگر بی‌نهایت باشد، مدل کامل (با حضور تمامی متغیرها) خواهد بود. $p(\boldsymbol{\beta})$

[11] Uniform kernel or boxcar kernel or rectangular kernel
[12] Triangular kernel
[13] Quartic kernel or biweight kernel
[14] Epanechnikov kernel
[15] Triweight kernel
[16] Tricube kernel
[17] Cosine kernel
[18] Laplacian of Gaussian kernel
[19] Gaussian kernel or normal kernel
[20] Tuning parameter

نیز تابعی از بردار پارامتر است که تابع تاوان[۲۱] نامیده شده و با توجه به نوع آن روش‌های تاوانیده متفاوتی ایجاد می‌شود. برای یافتن کمینه (بیشینه) یک تابع چند متغیره که با یک یا چند محدودیت مواجه است، روش لاگرانژ به کار گرفته می‌شود و با استفاده از روش لاگرانژ چند متغیره مسئله کمینه‌سازی (۶) به‌صورت زیر بازنویسی می‌شود [۱۳]

$$\min_{\boldsymbol{\beta}} \Big\{ (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})^\top (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) + \lambda p(\boldsymbol{\beta}) \Big\}. \qquad (۷)$$

در عبارت (۷) پارامتر نامنفی λ، پارامتر منظم‌سازی[۲۲] یا پارامتر تاوان نامیده می‌شود. اگر λ بی‌نهایت باشد، تنها عرض از مبدأ در مدل حضور خواهد داشت و اگر برابر صفر باشد، قسمت تاوان از مدل حذف‌شده و مدل کمترین توان‌های دوم حاصل خواهد شد.

پارامترهای λ و t نقش یکسانی در مدل ایفا می‌کنند و کنترل بزرگی یا کوچکی ناحیه محدودیت (تاوان) به عهده این دو پارامتر است. جالب‌توجه است که رابطه این دو پارامتر از لحاظ بزرگی معکوس است. لازم به ذکر است که در دو مسئله کمینه‌سازی (۶) و (۷) از تابع زیان توان دوم خطا[۲۳] استفاده شده است که بیانگر روش کمترین توان‌های دوم تاوانیده است؛ اما به‌طورکلی مسئله کمینه‌سازی تاوانیده به‌صورت زیر است:

$$\min_{\boldsymbol{\beta}} \Big\{ L(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}\boldsymbol{\beta}) + \lambda p(\boldsymbol{\beta}) \Big\},$$

که در آن $L(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}\boldsymbol{\beta})$ تابع زیان نامنفی برای نیکویی برازش، $p(\boldsymbol{\beta})$ تابع تاوان نامنفی و λ پارامتر منظم‌سازی است که تعادل بین نیکویی برازش و پیچیدگی مدل (حضور متغیرهای زیاد) را برقرار می‌کند [۱۷].

در ادامه به معرفی چندین روش تاوانیده پرداخته می‌شود.

۱.۳ رگرسیون لاسو

رگرسیون لاسو[۲۴] توسط رابرت تیبیشیرانی در سال ۱۹۹۶ ابداع شد. ایده اولیه لاسو برگرفته از پیشنهاد جالب لئو برایمن بود [۲۴]. روش گروتی برخی از متغیرها را حذف می‌کند و مابقی آن‌ها را منقبض می‌کند؛ بنابراین این روش مدل‌های قابل تفسیر ارائه می‌دهد. این روش نسبتاً پایدار است و همچنین مقیاس پایاست [۲۴]. وابستگی گروتی به علامت و مقدار برآورد کمترین توان‌های دوم نقص بزرگی است که بی‌اعتمادی محققان به این روش را به ارمغان آورده است. بدین‌جهت می‌توان گفت زمانی که با داده با بعد بالا مواجه باشیم و یا زمانی که متغیرها بسیار همبسته‌اند، برآوردهای کمترین توان‌های دوم ضعیف عمل کرده و به‌تبع آن گروتی نیز عملکرد ضعیفی داشته و مورداطمینان نیست. این در حالی است که لاسو برخلاف گروتی از استفاده مستقیم از برآوردگرهای کمترین توان‌های دوم اجتناب می‌کند. مسئله کمینه‌سازی لاسو به‌صورت

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=۱}^{n} \Big(\tilde{y}_i - \sum_{j=۱}^{p} \tilde{x}_{ij}\beta_j \Big)^۲ + \lambda \sum_{j=۱}^{p} |\beta_j| \right\},$$

است.

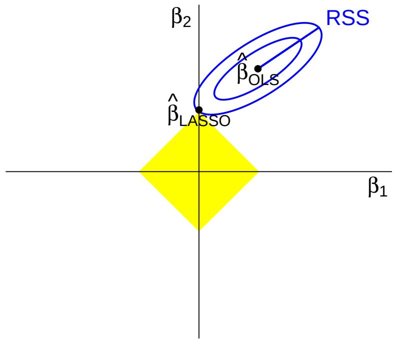
این محدودیت به‌طور طبیعی تمایل به تولید ضرایبی که صفر هستند، دارد و ازاین‌رو مدل‌های قابل تفسیر ارائه می‌دهد. روش لاسو برآوردگرهای پایدار، اریب و پیوسته تولید می‌کند [۲۴].

لاسو به معنی عملگر انتخاب و کمترین قدرمطلق انقباضی[۲۵] است. لاسو برخی ضرایب را منقبض و بقیه را صفر می‌کند.

برای یافتن درک بصری از روش لاسو با فرض وجود تنها دو متغیر توضیحی در مدل رگرسیونی، ناحیه محدودیت به‌صورت $|\beta_۱|+|\beta_۲| \le t$ یک لوزی است. در شکل ۱ لوزی زرد رنگ ناحیه تاوان است. محل برخورد کانتورهای RSS (مجموع توان‌های دوم باقی‌مانده‌ها) روش کمترین توان‌های دوم با لوزی، مختصات برآوردگر لاسو را نمایش می‌دهد. با عمود کردن از این نقطه بر محور افقی، $\hat{\beta}_۱$ و با عمود کردن از آن بر محور عمودی، $\hat{\beta}_۲$ حاصل می‌شود. اگر محل برخورد دقیقاً روی یکی از گوشه‌های لوزی باشد، (همانند شکل ۱) تنها یک متغیر انتخاب‌شده و دیگری صفر می‌شود. همان‌طور که ذکر شد، در روش لاسو به سبب نوع ناحیه تاوان (لوزی و گوشه داشتن لوزی) برخی از متغیرها دقیقاً برابر صفر می‌شوند. از منظر دیگر می‌توان استدلال کرد که چون تابع تاوان روش لاسو در نقطه صفر مشتق‌ناپذیر است، مقادیر بزرگ‌تری از λ منجر به حذف متغیرهای پیشگوی کم‌تأثیر در مدل می‌گردد و در نتیجه عمل برآورد پارامتر و انتخاب متغیر به‌صورت هم‌زمان انجام می‌گیرد [۲].

از معایب روش لاسو می‌توان به عملکرد ضعیف آن در حضور متغیرهای

---
[21] Penalty function
[22] Regularization parameter
[23] Square error loss function
[24] Lasso
[25] Least Absolute Shrinkage and Selection Operator (LASSO)



شکل ۱: ناحیه محدودیت لاسو.

توضیحی همبسته اشاره نمود. در این شرایط لاسو بین متغیرهای همبسته خصوصاً متغیرهایی که به‌صورت گروهی همبسته‌اند تنها یکی از آن‌ها را برمی‌گزیند و متغیر انتخاب‌شده لزوماً بهترین متغیر نیست. همچنین، لاسو برای یک مدل رگرسیونی خطی با p متغیر توضیحی و n مشاهده، حداکثر n متغیر را انتخاب می‌کند، بنابراین اگر متغیرهای بیشتری (بیشتر از n) در مدل معنی‌دار باشند، توسط لاسو انتخاب نمی‌شوند.

در روش کمترین توان‌های دوم پیراسته تعداد h تا از کوچک‌ترین توان‌های دوم باقی‌مانده‌های ($\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$) مسئله کمینه‌سازی مدل رگرسیونی خطی چندگانه به‌صورت

$$\min\left\{\sum_{i=1}^{h}(e^2)_{i:n}\right\}, \tag{۸}$$

در نظر گرفته می‌شود، به‌طوری‌که $(e^2)_{1:n} \le (e^2)_{2:n} \le \dots \le (e^2)_{n:n}$ توان دوم باقی‌مانده‌های مرتب‌شده هستند. لازم به ذکر است که ابتدا باقی‌مانده‌ها به توان دو می‌رسند، سپس مرتب می‌شوند. برخی h را پارامتر پیراسته[26] می‌نامند؛ اما برای تعیین این‌که مشاهده‌ی i ام یک مشاهده‌ی خوب است یا نه لازم است تابع نشانگر z_i را به‌صورت زیر تعریف نماییم.

$$z_i = \begin{cases} 1 & \text{مشاهده } i \text{ ام دورافتاده نباشد} \\ 0 & \text{مشاهده } i \text{ ام دورافتاده باشد} \end{cases}$$

و درنهایت ماتریس Z که یک ماتریس قطری است به‌صورت زیر حاصل می‌شود.

$$\boldsymbol{Z} = \begin{bmatrix} z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & z_n \end{bmatrix}.$$

بنابراین می‌توان مسئله کمینه‌سازی (۸) را به‌صورت

$$\begin{aligned} &\min_{\boldsymbol{\beta}}\left\{(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\beta})^\top \boldsymbol{Z}(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\beta})\right\}; \\ &\text{به‌طوری‌که} \quad \boldsymbol{e}^\top \boldsymbol{z} = h, \\ &z_i \in \{0,1\}, \quad i = 1,\dots,n, \end{aligned} \tag{۹}$$

بازنویسی کرد که در آن $\boldsymbol{z} = (z_1,\dots,z_n)^\top$ و $\boldsymbol{e} = (1,\dots,1)^\top_{n\times 1}$ است. با حل مسئله بهینه‌سازی (۹) برآوردگر کمترین توان‌های دوم پیراسته حاصل می‌شود. اگر $h = n$ باشد، کمترین توان‌های دوم پیراسته معادل کمترین توان‌های دوم معمولی خواهد بود.

یک نکته بحث‌برانگیز در این روش انتخاب پارامتر پیراسته است؛ زیرا h می‌تواند از $0,\dots,n$ باشد اگر این پارامتر به‌درستی انتخاب نشود، یا اطلاعات از دست می‌رود و یا نتایج حاصل تحریف می‌شود. روسو و لروی [۲۲] بابیان قضیه‌ای در کتاب خود نشان دادند که با انتخاب $[[\frac{n}{2}]]+[[\frac{p+1}{2}]]$ برای h، نقطه‌ی فروریزش (معیار استواری یک برآوردگر) این روش به ماکزیمم مقدار ممکن خود (۵۰ درصد) خواهد رسید؛ یعنی حتی اگر نیمی از مشاهدات پرت باشند، این روش همچنان استوار خواهد ماند؛ اما این امر همیشه و در هر شرایطی صادق نیست، به‌خصوص در داده با بعد بالا ضعف این قضیه به‌وضوح آشکار می‌شود.

[26] Trimmed parameter

یکی دیگر از معایب این روش یافتن h مشاهده عادی نیز امری بسیار دشوار و زمان‌بر است، زیرا باید $\binom{n}{h}$ مدل مختلف تشکیل شود و باهم مقایسه شوند. همین امر موجب پیچیدگی روش کمترین توان‌های دوم پیراسته و عدم استقبال محققان از آن تا قبل از تدوین بسته‌های نرم‌افزاری بود.

۴ روش کمترین توان‌های دوم پیراسته تنک

همان‌طور که ذکر شد روش لاسو تفسیر مدل‌های با بعد بالا را ساده می‌کند اما در حضور نقاط پرت مقاوم نیست و عملکرد ضعیفی دارد. به‌طور مثال نقطه فروریزش روش لاسو $\frac{1}{n}$ است؛ یعنی تنها حضور یک مشاهده پرت می‌تواند کلیه نتایج را زیر سؤال برده و بی‌اعتماد کند. علاوه بر این روش کمترین توان‌های دوم پیراسته در حالتی که $h < p$ باشد، نمی‌تواند محاسبه شود [۵].

روش‌های تاوانیده استوار، به‌عنوان روش‌های کمترین توان‌های دوم پیراسته تنک[27] نیز مشهور است. مسئله بهینه‌سازی این روش به‌صورت

$$\begin{aligned} &\min_{\beta} \left\{ (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})^\top \boldsymbol{Z}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) + h\lambda p(\boldsymbol{\beta}) \right\}; \\ &\text{به‌طوری‌که} \quad \boldsymbol{e}^\top \boldsymbol{z} = h, \\ &z_i \in \{0, 1\}, \quad i = 1, ..., n, \end{aligned} \tag{10}$$

خواهد بود به‌طوری‌که در آن $\boldsymbol{e} = (1, ..., 1)^\top_{n\times 1}$، $\boldsymbol{Z}$ یک ماتریس قطری با درایه‌های روی قطر اصلی $\boldsymbol{z} = (z_1, \ldots, z_n)^\top_{n\times 1}$، h پارامتر پیراسته و λ پارامتر تاوان است. انتخاب پارامتر پیراسته در ابعاد بالا امری به‌غایت دشوار است.

روش کمترین توان‌های دوم پیراسته مزایایی به شرح زیر داراست:

- عملکرد پیش‌بینی را از طریق کاهش واریانس بهبود می‌بخشد،
- به سبب انتخاب هم‌زمان، مدل تفسیرپذیری بیشتری را تضمین می‌کند،
- در حضور مجموعه داده با بعد بالا از مشکلات محاسباتی روش‌های کلاسیک جلوگیری می‌نماید.

۱.۴ نقطه فروریزش برآوردگر کمترین توان‌های دوم پیراسته

با استفاده از قضیه زیر که توسط آلفونز و همکاران [۵] مطرح شد، نقطه شکست یا نقطه فروریزش[28] برآوردگر کمترین توان‌های دوم پیراسته حاصل می‌شود.

قضیهٔ ۱.۴. فرض کنید $\rho(x)$ یک تابع زیان متقارن و محدب باشد، به‌طوری‌که $\rho(0) = 0$ و به ازای $x \neq 0$، $\rho(x) > 0$. همچنین $\boldsymbol{\rho}(\boldsymbol{x}) = (\rho(x_1), \ldots, \rho(x_n))^\top$ زیرمجموعه‌ای با اندازه $h \leq n$ و برآوردگر رگرسیونی

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \sum_{i=1}^{h} (\boldsymbol{\rho}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}))_{i:n} + h\lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

در نظر بگیرید، به‌طوری‌که

$$\left(\boldsymbol{\rho}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})\right)_{1:n} \leq \ldots \leq \left(\boldsymbol{\rho}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})\right)_{n:n}$$

توابع زیان مرتب‌شده هستند، در این صورت نقطه فروریزش برآوردگر رگرسیونی $\hat{\beta}$ به‌صورت

$$\varepsilon^*(\hat{\boldsymbol{\beta}}, Z^*) = \frac{n-h+1}{n},$$

است که در آن Z^* به‌عنوان تمامی نمونه‌هایی که دارای نقاط پرت و تأثیرگذار هستند، در نظر گرفته می‌شوند. این نمونه‌ها جایگزین هر زیرمجموعه‌ی m تایی از مشاهدات، در داده‌ها با مقادیر دلخواه می‌شوند.

با استفاده از قضیه ۱.۴ و با فرض این‌که $\rho(x) = x^2$ باشد، نقطه فروریزش برآوردگر کمترین توان‌های دوم پیراسته تنک برابر $\frac{n-h+1}{n}$ است. مقادیر کوچک h نقطه فروریزش بالاتر دارد و با فرض این‌که h به‌اندازه کافی کوچک باشد، نقطه فروریزش بزرگ‌تر از ۵۰٪ است [۵].

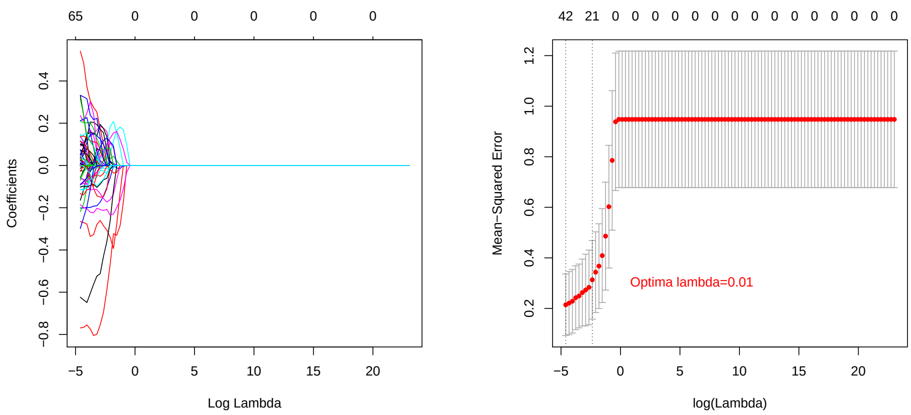
۵ نتایج کاربردی با داده ریبوفلاوین

در این بخش به‌منظور بررسی و تحلیل مدل‌های نیمه‌پارامتری پیراسته در ابعاد بالا از داده‌های ریبوفلاوین استفاده می‌شود. ریبوفلاوین که به‌عنوان ویتامین B2 نیز شناخته می‌شود، یکی از ویتامین‌های B است که همه محلول در آب هستند. از مهم‌ترین ویتامین‌های لازم برای حیات موجود زنده است. دلیل انتخاب این نام، رنگ زرد این ویتامین است که ناشی از وجود حلقه فلاوینی موجود در ساختمان آن است. ریبوفلاوین

[27] Sparse least trimmed squares
[28] Breakdown point

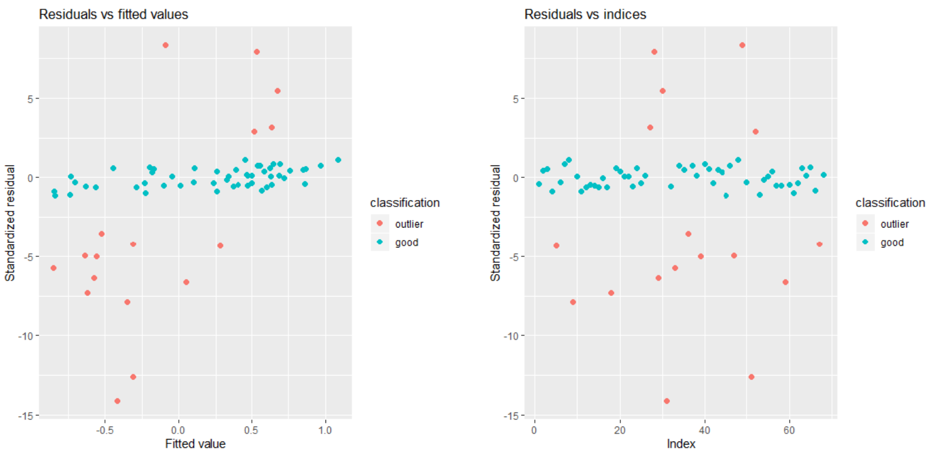
به‌طور طبیعی در برخی غذاها وجود دارد، به بعضی از محصولات غذایی اضافه می‌شود و به‌عنوان یک مکمل رژیم غذایی در دسترس است. این ویتامین یکی از اجزای اساسی دو کوآنزیم اصلی، فلاوین مونونوکلئوتید (FMN) و فلاوین آدنین دینوکلئوتید (FAD) است. این کوآنزیم ها نقش عمده‌ای در تولید انرژی، عملکرد سلولی، رشد و نمو و متابولیسم چربی‌ها، داروها و استروئیدها دارند. علاوه بر این، ریبوفلاوین به حفظ سطح طبیعی هموسیستئین، یک اسیدآمینه در خون کمک می‌کند. بیشترین درصد ریبوفلاوین پس از مصرف مواد گوشتی، سبزیجات، لبنیات و مقداری کمتر در مغزها و تخمه‌ها، حبوبات و سبوس غلات موجود توسط بدن دریافت می‌شود. بیشتر ریبوفلاوین توسط روده کوچک جذب و مقدار کمی در کبد، قلب و کلیه‌ها ذخیره‌شده و مابقی از طریق ادرار از بدن دفع خواهد شد. وضعیت ریبوفلاوین در افراد سالم به‌طورمعمول اندازه‌گیری نمی‌شود؛ اما در صورت بروز مشکل و تشخیص کمبود (وفور) این ویتامین از طریق بررسی گلبول‌های قرمز خون و یا بررسی میزان ریبوفلاوین دفع شده از طریق ادرار اندازه‌گیری خواهد شد. میزان میانگین ریبوفلاوین موردنیاز بدن روزانه بین ۱٫۱ تا ۱٫۳ میلی‌گرم برای بزرگ‌سالان و برای زنان باردار یا شیرده ۱٫۶ است. شایع‌ترین علت کمبود ریبوفلاوین در بدن رژیم غذایی نامناسب است. کمبود ریبوفلاوین همچنین می‌تواند در افرادی که دچار نقص در فعالیت‌های کبدی هستند ایجاد شود، چراکه مانع از استفاده مناسب از ویتامین‌ها می‌شود.

این داده‌ها در بسته نرم‌افزاری ”hdi“ نرم‌افزار R موجود است. در این مجموعه داده متغیر پاسخ لگاریتم نرخ تولید ریبوفلاوین است. این مجموعه داده دارای ۴۰۸۸ متغیر توضیحی بوده که هرکدام نشان‌دهنده لگاریتم سطح ژن‌ها است. با استفاده از روش تاوانیده لاسو استوار که در بسته نرم‌افزاری ”robustHD“ نرم‌افزار R موجود است، تعداد ۳۸ ژن به‌عنوان متغیرهای توضیحی مؤثر بر تغییرات و پیش‌بینی متغیر پاسخ بر اساس مقدار بهینه پارامتر تاوان ۰٫۰۰۴۰۸۷۱ شناسایی شدند. در شکل ۲ نمودارهای اعتبارسنجی متقابل ۱۰-سطحی و نمودار برآورد ضرایب به ازای مقادیر مختلف پارامتر تاوان نشان داده شده است. همچنین نمودار نقاط پرت شناسایی‌شده در شکل ۳ نمایش داده شده است.

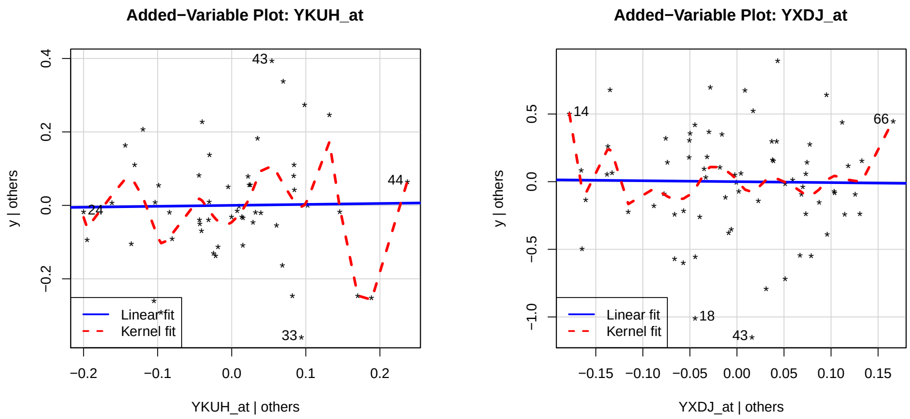


(ب) (الف)

شکل ۲: نمودارهای مربوط به پارامتر تاوان لاسو، الف: اعتبارسنجی متقابل ۱۰-سطحی، ب: برآورد ضرایب به ازای مقادیر مختلف پارامتر تاوان به روش لاسو.



شکل ۳: نمودار شناسایی نقاط پرت.



(ب) (الف)

شکل ۴: نمودار افزوده متغیرهای پیشگو دارای رابطه غیرخطی با متغیر پاسخ و برازش کم‌ترین توان‌های دوم (خط ممتد) و برازش ناپارامتری کرنل (خط‌چین)، الف: روش لاسو غیر استوار، ب: روش لاسو استوار

عنصر ناپارامتری مطابق [۸] با محاسبه

$$s_i^{۲} = \frac{۱}{n - p_۱ - ۱}\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}[, -i]\hat{\boldsymbol{\beta}}\right)^{\top}\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}[, -i]\hat{\boldsymbol{\beta}}\right),$$
$$i = ۱, \dots, ۴۰۸۸, \quad p_۱ = ۳۸,$$

شناسایی می‌شود که در آن $\tilde{\boldsymbol{y}}$ متغیر پاسخ (لگاریتم تولید ریبوفلاوین)، $\tilde{\boldsymbol{X}}$ ماتریس طرح و $\tilde{\boldsymbol{X}}[, -i]$ ماتریس طرح فراهم‌شده با حذف iامین ستون است. بدیهی است عنصر ناپارامتری، متغیر توضیحی است که با حذف آن میزان $s_i^{۲}$ کمینه خواهد شد. با استفاده از روش بالا به ترتیب متغیرهای توضیحی "$YXDJ\_at$" و "$YKUH\_at$" با استفاده از رویکردهای لاسو غیراستوار و لاسو استوار به‌عنوان عضو ناپارامتری شناسایی شدند. همچنین، با استفاده از نمودار متغیر افزوده[۲۹] نیز رابطه غیرخطی اجزای ناپارامتری مدل‌های لاسو غیراستوار و لاسو استوار مجدداً بطور شهودی در شکل ۴ بررسی و نتیجه به‌دست‌آمده در بالا تائید شد. نمودار متغیر افزوده بطور شهودی اثر هر یک از متغیرهای پیشگو را پس از حذف اثر سایر متغیرهای پیشگو، بر متغیر پاسخ آشکار می‌کند.

در جدول ۱ مقادیر برآوردگرها، انحراف از معیار و همچنین RSS، R^2 و CV که به ترتیب، مجموع توان‌های دوم مانده‌ها، ضریب تعیین و اعتبارسنجی متقابل مدل رگرسیون نیمه‌پارامتری برازش شده هستند، گزارش‌شده‌اند. همان‌طور که در این جدول دیده می‌شود، برآوردگرهای استوار لاسو دارای مجموع توان دوم خطا و اعتبارسنجی متقابل کمتری نسبت به برآوردگرهای غیر استوار لاسو بوده و حدود ۳۵ درصد بیشتر تغییرات متغیر پاسخ را نسبت به برآوردگر غیر استوار بیان می‌کنند.

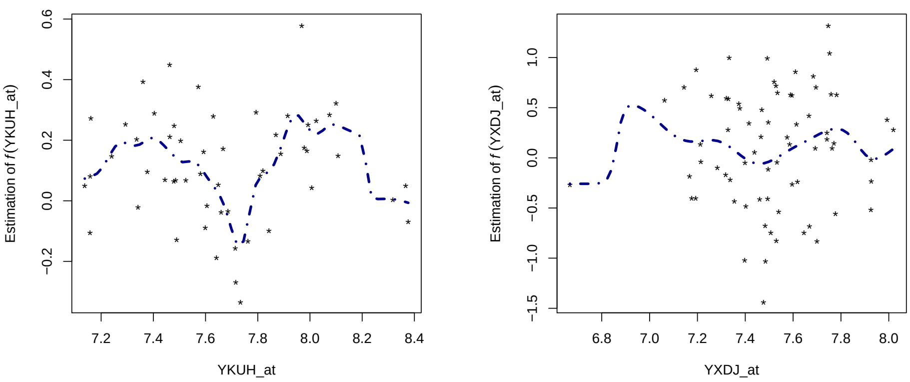
[29] Added Variable Plot

پس از برآورد ضرایب متغیرهای مؤثر شناسایی‌شده به روش تاوانیده لاسو غیراستوار و استوار، توابع ناپارامتری با استفاده از روش کرنل برازش گردید. مقدار بهینه پارامتر پهنای باند به روش کرنل به ترتیب برای متغیرهای توضیحی ناپارامتری "$YXDJat$" و "$YKUHat$" برابر ۰/۱۲۴۳ و ۰/۹۴۵ با رویکردهای لاسو غیراستوار و لاسو استوار تعیین شد و ۳۷ متغیر توضیحی دیگر به‌عنوان بخش خطی مدل در نظر گرفته شدند. توابع ناپارامتری برازش شده در شکل ۵ رسم شده است.

۶ بحث و نتیجه‌گیری

در این مقاله برآورد تاوانیده استوار لاسو برای ضرایب خطی و توابع غیرخطی در مدل رگرسیون نیمه‌پارامتری در حضور داده‌های پرت در داده‌های با بعد بالا به‌کمک الگوریتم سریع کم‌ترین توان‌های دوم پیراسته تنک، معرفی شد. نقطه فروریزش برآوردگر معرفی‌شده به دست آمد و از لحاظ نظری نشان داده شد که این نقطه فروریزش می‌تواند بیش از ۵۰ درصد باشد. بااین‌حال از دیدگاه عملی دستیابی به نقطه فروریزش بیش از ۵۰ درصد امکان‌پذیر نیست.

نتایج عددی مطالعه داده‌های واقعی نشان می‌دهد که استفاده از برآوردگر استوار لاسو تنک در مدل رگرسیون نیمه‌پارامتری، دارای تأثیر به سزایی در بهبود برآوردگرهای ضرایب خطی و توابع غیرخطی مدل در برازش مدل به داده‌های با بعد بالا و دارای مشاهدات دورافتاده است.



(ب) (الف)

شکل ۵: الف: برازش تابع ناپارامتری به روش غیراستوار لاسو ب: برازش تابع ناپارامتری به روش استوار لاسو

جدول ۱: ارزیابی برآوردگرهای پیشنهادی برای متغیرهای توضیحی مؤثر در داده‌های واقعی.

روش	لاسو غیراستوار		لاسو استوار	
	برآورد	انحراف استاندارد	برآورد	انحراف استاندارد
عرض از مبدأ	−۰٫۰۵۶۲۵	۰٫۰۳۳۹۸	−۰٫۰۶۵۲۶	۰٫۰۰۲۲۰
ALD_at	−۰٫۲۲۷۰۳	۰٫۱۶۰۴۳	۰٫۳۰۱۴۶	۰٫۰۲۴۱۵
$ALDX\_at$	−۰٫۰۹۱۷۲	۰٫۴۶۷۱۵	−۰٫۴۵۸۸۷	۰٫۰۷۳۲۷
$CSBA\_at$	۱٫۱۱۳۷۹	۰٫۳۷۰۵۳	۰٫۴۸۰۵۹	۰٫۰۳۴۶۳
$CTAA\_at$	−۰٫۶۸۲۵۶	۰٫۰۸۶۹۴	−۰٫۳۴۴۱۶	۰٫۰۱۴۰۱
$DEAD\_at$	−۰٫۱۷۵۳۷	۰٫۱۲۵۳۲	−۰٫۰۵۱۳۷	۰٫۰۱۳۹۶
$IOLJ\_at$	۰٫۰۷۹۰۱	۰٫۱۵۰۱۹	−۰٫۰۵۶۰۵	۰٫۰۱۶۹۹
$LACA\_at$	۰٫۴۱۲۰۱	۰٫۳۵۶۴۵	۰٫۴۵۸۸۸	۰٫۰۵۸۶۷
$MMGE\_at$	۰٫۶۹۶۳۴	۰٫۳۷۰۳۷	−۰٫۱۳۵۱۴	۰٫۰۳۱۹۳
$MREBH\_at$	−۰٫۴۲۵۳۱	۰٫۴۵۸۰۷	۰٫۰۲۴۰۸	۰٫۰۷۳۰۹
$PCKA\_at$	۱٫۵۲۳۲۹	۰٫۳۸۲۳۶	−۰٫۲۶۱۳۶	۰٫۰۴۹۶۵
$PKSI\_at$	−۰٫۳۲۲۷۲	۰٫۳۵۴۷۳	۰٫۳۶۴۳۰	۰٫۰۳۶۲۰
$PKSM\_at$	۰٫۵۰۹۵۱	۰٫۶۰۶۳۱	−۰٫۰۵۰۱۲	۰٫۰۷۶۲۱
PTA_at	۱٫۶۷۶۶۱	۰٫۶۲۴۴۵	۱٫۲۶۹۰۰	۰٫۱۵۳۰۳
$RIBT\_at$	−۰٫۳۶۶۱۵	۰٫۴۴۲۹۶	−۰٫۳۵۱۴۵	۰٫۰۶۰۲۹
$SPOIISA\_at$	−۱٫۲۴۱۸۹	۰٫۳۴۱۱۲	−۰٫۶۶۷۰۱	۰٫۱۴۰۶۸
$SPOIISB\_at$	−۰٫۷۲۴۴۷	۰٫۶۹۳۵۸	۰٫۰۸۶۷۲	۰٫۰۴۸۵۱
$XKDS\_at$	۰٫۵۴۰۶۸	۰٫۰۸۹۰۳	−۰٫۰۵۹۵۴	۰٫۰۱۰۴۲
$YCEI\_at$	−۰٫۱۲۵۰۷	۰٫۲۳۱۲۵	۰٫۵۱۲۸۹	۰٫۰۲۳۱۹
$YCGR\_at$	−۰٫۰۷۲۲۷	۰٫۳۵۱۷۷	۰٫۰۷۲۰۹	۰٫۰۵۴۰۰
$YDBM\_at$	−۰٫۳۶۴۱۴	۰٫۴۳۵۳۸	−۰٫۹۹۰۸۷	۰٫۰۶۷۹۴
$YEBC\_at$	−۰٫۳۲۷۰۰	۰٫۴۵۱۵۷	−۰٫۱۰۰۲۲	۰٫۰۸۴۰۰
$YFHE\_r\_at$	۰٫۹۲۸۲۹	۰٫۳۷۲۲۶	۱٫۱۴۶۷۹	۰٫۱۰۵۴۱
$YHCL\_at$	۰٫۰۲۱۷۱	۰٫۳۷۶۶۶	−۱٫۰۶۴۵۸	۰٫۰۵۹۵۱
$YJBJ\_at$	−۰٫۲۷۱۵۰	۰٫۶۴۵۹۵	۰٫۶۵۲۲۲	۰٫۰۵۱۳۳
$YJCJ\_at$	۱٫۸۴۶۰۶	۰٫۴۹۵۵۳	−۰٫۲۷۵۱۷	۰٫۰۴۲۱۳
$YKUH\_at$	−۱٫۳۶۷۹۰	۰٫۳۱۶۶۵	−۰٫۳۵۶۳۰	۰٫۰۳۱۹۲
$YORB\_i\_at$	۰٫۲۶۲۷۰	۰٫۲۲۵۶۳	−۰٫۴۴۷۱۰	۰٫۱۶۶۱۰
$YQET\_at$	−۲٫۶۲۲۷۴	۰٫۶۲۰۷۵	−۰٫۴۸۷۷۲	۰٫۰۱۴۷۹
$YQHJ\_at$	−۰٫۱۰۹۶۱	۰٫۰۹۶۶۹	−۰٫۲۷۹۱۲	۰٫۰۱۷۶۵
$YUNL\_at$	−۰٫۶۴۴۳۵	۰٫۱۹۳۵۲	۰٫۵۹۳۶۴	۰٫۰۲۳۴۶
$YWFM\_at$	۰٫۴۴۰۲۵	۰٫۲۲۳۲۳	۰٫۴۹۰۱۱	۰٫۰۲۱۰۹
$YWPB\_at$	−۰٫۲۸۱۲۶	۰٫۱۷۷۶۶	−۱٫۰۰۰۳۵	۰٫۰۷۸۸۲
$YXDJ\_at$	۰٫۰۶۲۷۹	۰٫۰۵۶۲۹	۰٫۲۷۳۸۹	۰٫۰۰۵۴۲
$YXER\_at$	−۱٫۸۴۹۷۹	۰٫۳۶۱۳۶	−۰٫۸۴۵۲۷	۰٫۰۵۵۶۲
$YXJL\_at$	−۰٫۱۱۳۸۴	۰٫۱۴۹۷۱	−۰٫۰۲۳۱۸	۰٫۰۲۹۵۶
$YXLD\_at$	۰٫۶۳۰۱۴	۰٫۴۱۲۵۲	۰٫۱۲۰۶۷	۰٫۰۷۸۲۱
$YXLE\_at$	۰٫۰۸۷۹۸	۰٫۲۰۹۰۸	۰٫۴۳۸۳۱	۰٫۰۱۴۰۹
RSS	۲۷٫۲۵۱۳		۲٫۰۰۱۳	
R^2	۰٫۵۴۰۵		۰٫۸۹۲۸	
CV	۰٫۲۷۸۰		۰٫۱۶۱۸	

مراجع

[۱] آرشی، م.، صادقی، ح؛ و طباطبایی م. (۱۳۹۸)، استنباط آماری، انتشارات دانشگاه صنعتی شاهرود، شاهرود.

[۲] جذن،س؛ و امینی، م. (۱۳۹۶)، برآورداستوار نسبت به مشاهده‌های دورافتاده در رگرسیون خطی در حضور هم‌خطی چندگانه، *اندیشه آماری*، ۴۴، ۹۳-۱۱۰.

[۳] روزبه، م؛ و چاچی، ج. (۱۳۹۵)، برآورد تفاضلی استوار مدل‌های خطی جزئی، *مجله علوم آماری*، ۱۰(۱)، ۹۵-۱۱۲.

[4] Akdeniz, D.E., Hardle, W.K., and Osipenko, M. (2012), Difference Based Ridge and Liu Type Estimators in Semiparametric Regression Models, *Journal of Multivariate Analysis*. **105(1)**, 164-175.

[5] Alfons, A., Croux, C., and Gelper, S. (2013), Sparse Least Trimmed Squarse Regression for Analyzing High-Dimesional Large Data Set, *The Annals of Applied Statistics*. **7(1)**, 226-248.

[6] Amini, M., and Roozbeh, M. (2015), Optimal Partial Ridge Estimation in Restricted Semiparametric Regression Models, *Journal of Multivariate Analysis*. **136(2)**, 26-40.

[7] Arashi, M., Asar, Y., and Yuzbasi B. (2021), SLASSO: a scaled LASSO for multicollinear situations, *Journal of Statistical Computation and Simulation*. **91(15)**, 3170-3183.

[8] Arashi, M., and Roozbeh, M. (2019) , Some Improved Estimation Strategies in High-dimensional Semiparametric Regression Models with Application to Riboflavin Production Data, *Journal of Statistical Papers*, **60**, 667-686.

[9] Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data*. JohnWiley ad Sons, New York.

[10] Bunea, F. (2004), Consistent Covariate Selection and Post Model Selection Inference in Semiparametric Regression, *The Annals of Mathematical Statistics*. **32(3)**, 898-927.

[11] Efron,B., and Haste.T. (2017), *Computer Age Statistical Inference*. Cambridge University Press, Cambridge.

[12] Engle, R. F., Granger, C.W.J., Rice, J., and Weiss, A. (1986), Semiparametric Estimation of the Relation Between Weather and Electricity Sale, *Journal of the American statistical Association*. **81(394)**, 310-320.

[13] Everitt, B., and Hothorn, T. (2011), *An Introduction to Applied Multivariate Analysis with R*. Springer, New York Dordrecht, Heidelberg, London.

[14] Fan, J., Hardle, W., and Mammen, E. (1998), Direct Estimation of Low Dimensional Components in Additive Models, *The Annals of Statistics*. **26(3)**, 943-971.

[15] Gannaz, I. (2007), Robust Estimation and Wavelet Thresholding in Partially Linear Models, *Statistics and Computing*. **17(4)**, 293-310.

[16] Hardle, W.k., Laing, H., and Gao, J. (2000), *Partially Linear Models*. PhysikaVerlag, Heidelberg.

[17] Li, B., and Yu, Q. (2009), Robust and Sparse Bridge Regression, *Statistics and its interface*, **2(4)**, 481-491.

[18] Nadaraya, E.A. (1964), On Estimating Regression, *Theory of Probability and its Applications*. **9(1)**, 141-142.

[19] Roozbeh, M. (2015), Shrinkage Ridge Estimators in Semiparametric Regression Models, *Journal of Multivariate Analysis*. **136(2)**, 56-74.

[20] Roozbeh, M. (2018), Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion, 117, *Computational Statistics & Data Analysis*. **117**, 45-61.

[21] Roozbeh, M., and Arashi, M. (2013), Feasible Ridge Estimator in Partially Linear Models, *Journal of Multivariate Analysis*. **116(4)**, 35-44.

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

[23] Speakman, P. (1988), Kernel Somoothing in Partial Linear Models, *Journal of the Royal Statistical Society: Series B (Methodological)*. **50(1)**, 413-436.

[24] Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*. **58(1)**, 267-288.

[25] Yuzbasi B., Arashi, M., and Akdeniz, F. (2021), Penalized regression via the restricted bridge estimator, *Soft Computing*. **25(13)**, 8401-8416

[26] Wasserman, L. (2006), *All of Nonparametric statistics*. Springer Science and Business Media, New York.

[27] Watson, G. S. (1964), Smooth Regression Analysis, *Sankhyā: The Indian Journal of Statistics, Series A*. **26(4)**, 359-372.

Sparse robust semiparametric models in high-dimensional data

Mahdi Roozbeh[1], Monireh Manavi[2]

Abstract:

Analysis and modeling the high-dimensional data is one of the most challenging problems faced by the world nowaday. Interpretation of such data is not easy and needs to be applied to modern methods. The penalized methods are one of the most popular ways to analyze the high-dimensional data. Also, the regression models and their analysis are affected by the outliers seriously. The least trimmed squares method is one of the best robust approaches to solve the corruptive influence of the outliers. Semiparametric models, which are a combination of both parametric and nonparametric models, are very flexible models. They are useful when the model contains both parametric and nonparametric parts. The main purpose of this paper is to analyze semiparametric models in high-dimensional data with the presence of outliers using the robust sparse Lasso approach. Finally, the performance of the proposed estimator is examined using a real data analysis about production of vitamin B2.

Keywords: High-dimensional data, Lasso method, Least trimmed squares method, Semiparametric model, Sparse least trimmed squares method.

[1]**Faculty of mathematics, Semnan university, Semnan, Iran.**

[2]**Master's degree graduate, statistics and Computer science, Semnan university, Semnan, Iran.**