کاربرد رگرسیون ستیغی کمترین توان‌های دوم پیراسته محدودشده تصادفی در مدل‌سازی مصرف آب

مهدی روزبه[۱]، ملیحه سادات ملک‌جعفریان[۲] و منیره معنوی[۳]

تاریخ دریافت: ۱۴۰۰/۰۲/۳۱
تاریخ پذیرش: ۱۴۰۰/۱۲/۲۶

چکیده:
مهم‌ترین هدف علم آمار تجزیه‌وتحلیل داده‌های واقعی دنیای پیرامون بشر است. اگر این اطلاعات دقیق و درست تحلیل شوند، نتایج حاصل در بسیاری از تصمیمات مهم یا ریگر ما خواهد بود. از جمله داده‌های واقعی پیرامون ما که تحلیل آن بسیار مهم است، داده‌های مربوط به مصرف آب می‌باشد. با توجه به اینکه کشور ایران در ناحیه نیمه‌خشک آب و هوایی از کره زمین قرار دارد، لازم است برای پیش‌بینی و برگزیدن بهترین و مناسب‌ترین مدل‌های دقیق مصرف آب گام‌های ژرفی برداشت که لازمه تصمیمات کلان کشوری می‌باشد. در تجزیه‌وتحلیل داده‌های واقعی ممکن است محقق با مشکل هم خطی و نقاط دورافتاده مواجه شود. روش‌های مقاوم (استوار) برای تحلیل مجموعه داده‌های دارای نقاط دورافتاده و رویکرد ستیغی روشی است که برای تحلیل مجموعه داده‌های دارای هم خطی استفاده می‌شوند. محدودیت روی مدل‌ها نیز ناشی از به‌کارگیری اطلاعات غیرنمونه‌ای در برآورد ضرایب رگرسیونی است. در این مقاله به مدل‌سازی داده‌های مصرف آب، با استفاده از رویکرد ستیغی محدودشده تصادفی استوار پرداخته می‌شود.

واژه‌های کلیدی: روش کمترین توان‌های دوم پیراسته ستیغی، محدودیت خطی تصادفی، مصرف آب، نقاط دورافتاده، هم خطی.

۱ مقدمه

تجزیه‌وتحلیل درست این داده‌های واقعی دنیای پیرامون بشر اطلاعات بسیار مهمی در مورد جامعه در اختیار محقق قرار می‌دهد. این اطلاعات و پیش‌بینی‌های حاصل از آن در تصمیم‌گیری‌های مهم و حیاتی در دنیای واقعی به کمک مسئولین خواهند آمد. اگر کوچک‌ترین اشتباهی در این تحلیل‌ها رخ دهد نتایج نامطلوب آن تا مدت‌ها گریبان‌گیر جامعه خواهد بود. بنابراین می‌بایست تحلیل‌ها با بیشترین دقت ممکن انجام شوند. اما حقیقت این است که داده‌های واقعی موجود در جامعه دارای پیچیدگی خاصی هستند و به‌سادگی و با روش‌های ساده موجود قابل تجزیه‌وتحلیل نیستند. یکی از رایج‌ترین مشکلات موجود در داده‌های واقعی هم خطی[۴] است. یک مسئله جدی که می‌تواند استفاده از مدل رگرسیونی را با اشکال مواجه کند، هم خطی چندگانه یا وابستگی خطی نزدیک بین متغیرهای رگرسیونی است. وجود وابستگی خطی نزدیک، توانایی برآورد ضرایب مدل رگرسیونی با استفاده از روش کمترین توان‌های دوم معمولی را با مشکل مواجه می‌کند. در بیشتر کاربردهای رگرسیون، بین متغیرهای رگرسیونی مسئله هم خطی چندگانه وجود دارد و استنباط‌هایی نظیر مشخص کردن اثرات نسبی متغیرهای رگرسیونی، پیش‌گویی و برآورد و انتخاب یک مجموعه مناسب از متغیرها بر اساس این الگوی رگرسیون ممکن است تا حد زیادی گمراه‌کننده و پراشتباه باشد.

منابع اصلی هم خطی به‌صورت زیر است [۱۲]:

- شیوه جمع‌آوری داده‌ها.
- گذاشتن قیدهایی روی مدل.
- در نظر گرفتن یک مدل با متغیرهای رگرسیونی بیش‌ازحد نیاز.

[۱] هیئت‌علمی گروه آمار، دانشگاه سمنان، سمنان، ایران (نویسنده مسئول: mahdi.roozbeh@semnan.ac.ir)
[۲] شرکت آب و فاضلاب شهری استان سمنان، سمنان، ایران
[۳] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.

[4] Collinearity

وجود هم خطی ممکن است منجر به فواصل اطمینان عریض برای پارامترها، ناپایداری برآورد پارامترها[5] یا تولید برآوردهایی با علامت اشتباه شود [۱]. روش‌های مختلفی برای غلبه برای هم خطی وجود دارد از جمله رگرسیون مؤلفه‌ی اصلی[6]، رگرسیون ستیغی[7] و ... .

مفهوم رگرسیون ستیغی توسط هورل و کنارد [۹] برای مقابله با هم خطی در مسائل رگرسیون پیشنهادشده است. از آن زمان تاکنون بسیاری از نویسندگان و محققان از این روش برای تحقیقات و پژوهش‌های خود استفاده کرده‌اند.

اولین و مهم‌ترین گام جهت انجام تحقیق، جمع‌آوری داده‌ها است. اغلب در بین مجموعه داده‌ها چند مشاهده وجود دارد که با سایر مشاهدات تفاوت دارند که تحت نام نقاط دورافتاده[8] یاد می‌شوند. داده‌های دورافتاده مقادیر بسیاری از آماره‌ها و برآوردگرها از جمله میانگین و واریانس را تحت تأثیر قرار داده و موجب نتایج اشتباه می‌گردند. به‌عنوان‌مثال این داده‌ها با اثرگذاری بر برآورد ضرایب مدل‌های رگرسیون، مجموع توان‌های دوم خطا و غیره سبب ارائه نتایج غیرواقعی می‌گردند. حذف این‌گونه داده‌ها سبب از بین رفتن اطلاعات می‌شود. بنابراین بهترین راه برخورد با داده دورافتاده، استفاده از روش‌های استوار[9] است به‌طوری‌که آماره‌ها و برآوردگرهای این روش تحت تأثیر داده دورافتاده قرار نگیرد.

زمانی که یک مشاهده تفاوت اساسی با دیگر مشاهدات داشته باشد، می‌تواند یک انحراف اساسی در نتایج تحلیل رگرسیونی ایجاد کند [۱۱]. محققان در مواجهه با این نقاط علاقه‌مند هستند که این مجموعه از داده‌ها را، که به آن‌ها نقاط دورافتاده نیز گفته می‌شود، تعیین و تأثیر آن‌ها را بر جنبه‌های مختلف تحلیل رگرسیونی ارزیابی کنند. نقاط دورافتاده ممکن است نتایج برازش مدل را تحت تأثیر قرار دهند، به‌گونه‌ای که حذف آن‌ها از مجموعه داده‌ها نتایج کاملاً متفاوتی به بار آورد. روش‌ها و معیارهای مختلفی به‌منظور شناسایی چنین مشاهداتی پیشنهادشده است. البته برخی روش‌های استوار، به مشاهدات دورافتاده وزن صفر را اختصاص می‌دهند و عملاً مشاهده دورافتاده از مجموعه داده‌ها حذف می‌شود اما روش‌های دیگری نیز وجود دارد که به مشاهدات دورافتاده وزن غیرصفر ولی ناچیز اختصاص می‌دهند ([۱۴]، [۱۶] و [۱۷]).

در شکل ۱ ملاحظه می‌گردد وجود داده دورافتاده، برازش خط رگرسیون را تحت تأثیر قرار داده به‌طوری‌که با حذف این نقاط خط رگرسیونی به‌شدت تغییر می‌کند.

روش‌های متنوع و بسیاری برای شناسایی نقاط دورافتاده، نظیر تحلیل باقی‌مانده‌ها[10] و ماتریس کلاهدار[11]، پتانسیل‌ها[12]، آماره نسبت کوواریانس[13] و... وجود دارد. به‌طورکلی می‌توان روش‌های شناسایی نقاط را به دو گروه روش‌های عددی و روش‌های شهودی نیز تقسیم نمود که این روش‌ها به‌طور کامل در [۲] و [۶] قابل‌مشاهده است.

یکی دیگر از نکات بسیار مهم در برآورد ضرایب رگرسیونی یافتن محدودیت‌های موجود روی فضای پارامتر و توجه به آن‌ها در برآورد است. زیرا این محدودیت‌ها روی مقادیر به‌دست‌آمده و برآورد ضرایب تأثیر زیادی خواهند داشت. کاملاً واضح است که محدودیت فضای پارامتر فضای برآورد را نیز محدود خواهد کرد و می‌بایست برآوردهای به‌دست‌آمده نیز در این فضای محدودشده صدق کنند و نمی‌توان برآوردی یافت که در فضای محدودشده نباشند. بنابراین نادیده گرفتن این محدودیت‌ها، درستی برآوردهای به‌دست‌آمده را تحت‌الشعاع قرار می‌دهد. بدین‌جهت می‌بایست ضرایب رگرسیونی، تحت این محدودها روی مدل اعمال شوند [۱۵].

مدل‌های محدودشده به‌طور گسترده‌ای در مسئله آزمون فرضیات آماری، به‌ویژه آزمون نسبت درست نمایی تعمیم‌یافته در مدل‌های رگرسیون کاربرد دارند. محدودیت روی پارامترهای مدل می‌تواند به یکی از دلایل زیر رخ دهد:

۱. حقیقت ناشی از ملاحظات نظری و یا تجربی (آزمایشگاهی)،
۲. فرضیه‌ای که باید آزمون شود،
۳. نظر محقق یا متخصص،
۴. شرط غیرمعمول به‌منظور کاهش یا حذف فزونگی روی مدل موردنظر [۲۰].

در دو بخش بعدی به معرفی دو روش کمترین توان‌های دوم معمولی و روش ستیغی می‌پردازیم.

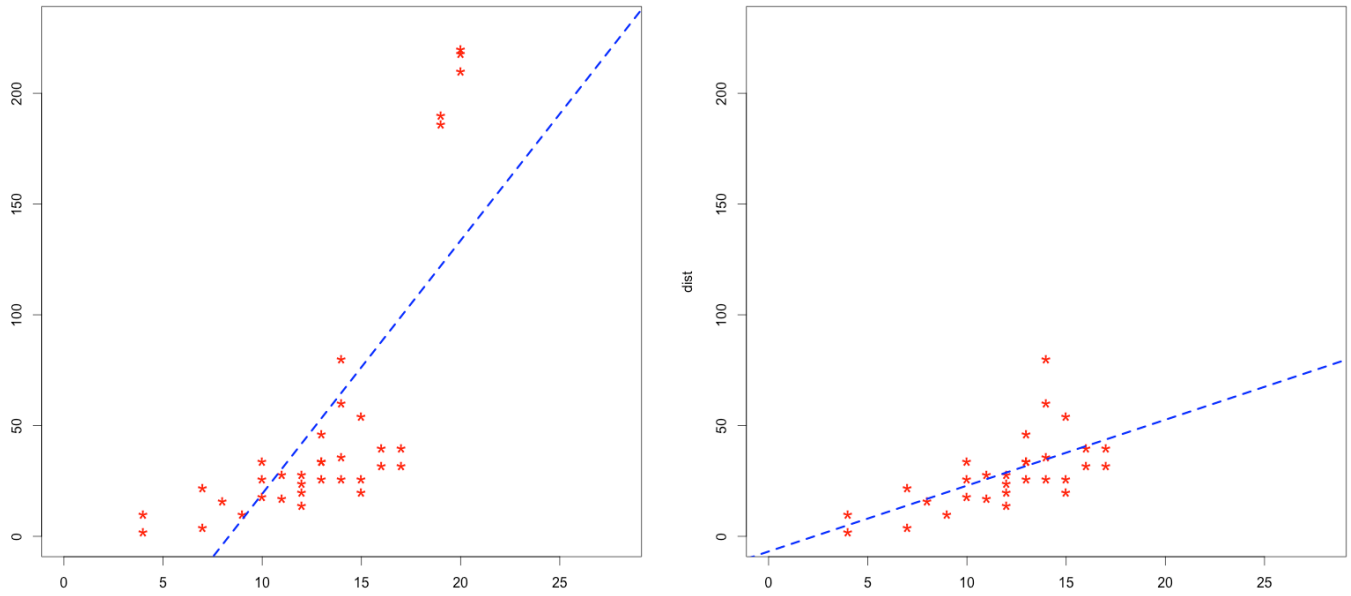
[5] Instability of parameters estimation
[6] Principal component regression (PCR)
[7] Ridge regression
[8] Outlier
[9] Robust methods
[10] Analysis of residuals
[11] Hat matrix
[12] Potentials
[13] Covariane ratio statistics

۱.۱ روش کمترین توان‌های دوم معمولی

مدل رگرسیونی خطی چندگانه را به‌صورت

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

تعریف می‌شود که در آن $\mathbf{y} = (y_1, \ldots, y_n)^\top$ متغیر پاسخ، $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top, j = 1, \ldots, p$ ستون j ام ماتریس $\mathbf{X}$ یعنی متغیر توضیحی j ام، $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ بردار ضرایب رگرسیونی و $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ بردار خطای تصادفی با شرایط $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ و $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \mathbf{I}_n$ است. لازم به ذکر است که منظور از 0 بردار صفر با n مؤلفه صفر و $\mathbf{I}_n$ ماتریس همانی $n \times n$ است.



شکل ۱: نمودار سمت چپ نمایش تأثیر دادهای دورافتاده، نمودار سمت راست حذف دادهای دورافتاده و برازش مجدد خط رگرسیون

برآورد ضرایب رگرسیونی با استفاده از روش کمترین با حل مسئله رگرسیونی

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

محاسبه می‌شود. طبق قضیه گاوس مارکوف برآوردگر به‌دست‌آمده با این روش، یعنی $\hat{\beta}_{\mathrm{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ بهترین برآوردگر خطی نااریب با کمترین واریانس است.

۲.۱ روش ستیغی

در صورت وجود هم خطی از روش ستیغی برای برآورد ضرایب رگرسیونی استفاده می‌شود که این برآوردگر با حل مسئله

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + k\boldsymbol{\beta}^\top \boldsymbol{\beta},$$

حاصل می‌شود. برآوردگر روش ستیغی، یعنی $\hat{\beta}_{\mathrm{Ridge}} = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$ یک برآوردگر اریب است.

۲ مدل‌های محدودشده

مدل‌های محدودشده، به دو گروه تصادفی و غیر تصادفی تقسیم می‌شوند که در ادامه معرفی می‌شوند.

تعریف ۱.۲. *محدودیت خطی غیر تصادفی*[14]*: اگر محدودیت به‌صورت*

$$\dot{\mathbf{r}} = \dot{\mathbf{R}}\boldsymbol{\beta}, \qquad (۲)$$

بیان شود، محدودیت خطی غیر تصادفی نام دارد به‌طوری‌که در آن $\dot{\mathbf{r}}$ بردار معلوم $q \times 1$ و $\dot{\mathbf{R}}$ ماتریس $q \times p$ اطلاعات پیشین معلوم روی بردار پارامترها است. ماتریس $\dot{\mathbf{R}}$ رتبه کامل سطری است.

لازم به ذکر است که منظور از اطلاعات پیشین، اطلاعات غیر نمونه‌ای در مورد پارامتر مجهول است به بیان ساده‌تر خلاصه‌ای از اطلاعات و دانسته‌های کاربر آمار در مورد پارامتر مجهول است.

[14] Linear restricted

طبق [۱۵] با ترکیب محدودیت خطی غیر تصادفی (۲) و مدل (۱) برآوردگر محدودشده غیر تصادفی به‌صورت زیر قابل‌محاسبه است:

$$\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}}_{ols} + (\mathbf{X}^\top\mathbf{X})^{-1}\dot{\mathbf{R}}(\dot{\mathbf{R}}(\mathbf{X}^\top\mathbf{X})^{-1}\dot{\mathbf{R}}^\top)^{-1}(\dot{\mathbf{r}} - \dot{\mathbf{R}}\hat{\boldsymbol{\beta}}_{ols}),$$

که در آن $\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ است.

سرکار [۱۹] برآوردگر ستیغی محدودشده غیر تصادفی را برای مواردی که هم خطی در مجموعه داده‌ها وجود داشته باشد، به‌صورت $\hat{\boldsymbol{\beta}}^{*R}_{Ridge} = (\mathbf{I}_p + k(\mathbf{X}^\top\mathbf{X})^{-1})^{-1}\hat{\boldsymbol{\beta}}^R$ معرفی کرد که در آن k پارامتر ستیغی نام دارد. متأسفانه عبارت $\dot{\mathbf{R}}\hat{\boldsymbol{\beta}}^{*R}_{Ridge} = \dot{\mathbf{r}}$ برای این برآوردگر برای هر $\boldsymbol{\beta}$ برقرار نخواهد بود [۷]. بنابراین برآوردگر $\hat{\boldsymbol{\beta}}^{*R}_{Ridge}$ برآوردگر مناسبی نبوده و لذا او برآوردگر

$$\hat{\boldsymbol{\beta}}^R_{Ridge} = \hat{\boldsymbol{\beta}}_{Ridge} + (\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1} \times(\dot{\mathbf{R}}^\top(\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}\dot{\mathbf{R}})^{-1}(\dot{\mathbf{r}} - \dot{\mathbf{R}}\hat{\boldsymbol{\beta}}_{Ridge}), \quad (۳)$$

را پیشنهاد کرد به‌طوری‌که $\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ است.

۱.۲ تعیین پارامتر ستیغی

تعیین مقدار بهینه پارامتر ستیغی نقش بسیار مهمی در نتایج نهایی دارد. مقدار بهینه این پارامتر توازن بین اریبی و واریانس را برقرار می‌کند. در این شرایط اریبی و واریانس به‌طور هم‌زمان به کمترین حالت ممکن خود می‌رسند. برای محاسبه این پارامتر از روش‌های مختلفی می‌توان استفاده کرد که در این مقاله از روش اعتبارسنجی متقابل[۱۵] استفاده می‌شود. اعتبارسنجی متقابل انواع مختلفی دارد که در اینجا روش حذف تکی[۱۶] مورداستفاده قرار می‌گیرد. این روش برای مدل رگرسیونی (۱) به‌صورت

$$\text{LOCV} = \|\mathbf{y} - \hat{\mathbf{y}}_{(-i)}\|^2,$$

تعریف می‌شود که در آن $\hat{\mathbf{y}}_{(-i)} = \mathbf{X}_{(-i)}\hat{\boldsymbol{\beta}}_{(-i)}$ به‌طوری‌که $\hat{\boldsymbol{\beta}}_{(-i)}$ برآوردگر دلخواه است که با حذف i امین سطر ماتریس $\mathbf{X}$ $(\mathbf{X}_{(-i)})$ و i امین مشاهده بردار $\mathbf{y}$ $(\mathbf{y}_{-i})$ به دست می‌آید. به بیان ساده‌تر یعنی مدلی با حذف مشاهده i ام برازش داده می‌شود و بعد با استفاده از مدل برازش داده‌شده مشاهده y_i پیش‌بینی می‌شود. به‌عبارت‌دیگر $\hat{y}_i$ مقدار پیش‌بینی‌شده مشاهده iام از مدل حاصل با حذف مشاهده iام است.

در عمل، موقعیت‌های بسیاری وجود دارد که در آن فرض می‌شود که اطلاعات پیشین مناسب نیستند و یا به بیان ساده‌تر، در حقیقت اطلاعات پیشین، دقیق نبوده و می‌بایست محدودیت به‌صورت تصادفی بیان می‌شود [۵]. محدودیت خطی تصادفی در مواردی نظیر روابط اقتصادی، ساختارهای صنعتی، برنامه‌ریزی‌های تولیدی و ... کاربرد دارند [۴]. اگر عدم اطمینان در مورد اطلاعات پیشین وجود داشته باشد، یک جایگزین مناسب برای محدودیت‌های خطی، محدودیت‌های خطی تصادفی است.

تعریف ۲.۲. *محدودیت خطی تصادفی[۱۷]: اگر محدودیت به‌صورت*

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\nu}, \quad (۴)$$

بیان شود، محدودیت خطی تصادفی نام دارد به‌طوری‌که در آن $\mathbf{r}$ بردار معلوم $q \times 1$، $\mathbf{R}$ ماتریس $q \times p$ اطلاعات پیشین معلوم روی بردار پارامترها و $\boldsymbol{\nu}$ خطای تصادفی (اطلاعات پیشین مجهول) است. ماتریس $\mathbf{R}$ سطری رتبه کامل است.

تیل و گلدبرگر [۲۱] با اعمال محدودیت خطی تصادفی (۴) در مدل (۱)، مدل جدیدی با فرض استقلال $\boldsymbol{\varepsilon}$ و $\boldsymbol{\nu}$ به‌صورت

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{r}\end{bmatrix} = \begin{bmatrix}\mathbf{X}\\ \mathbf{R}\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\nu}\end{bmatrix},$$

به دست آوردند که می‌توان آن را به‌صورت

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\varepsilon}, \quad \tilde{\mathbf{y}} = \begin{bmatrix}\mathbf{y}\\ \mathbf{r}\end{bmatrix}, \ \tilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X}\\ \mathbf{R}\end{bmatrix}, \ \tilde{\varepsilon} = \begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\nu}\end{bmatrix}, \quad (۵)$$

بازنویسی کرد که در واقع، همان مدل رگرسیونی خطی ساده با شرایط $\mathrm{E}(\tilde{\varepsilon}) = 0$ و $\mathrm{E}(\tilde{\varepsilon}\tilde{\varepsilon}^\top) = \sigma^2\mathbf{I}_{(n+q)\times p}$ است.

برآورد ضرایب رگرسیونی مدل محدودشده تصادفی با فرض وجود هم خطی در مجموعه داده‌ها با روش ستیغی نیز به‌صورت

$$\min_{\boldsymbol{\beta}}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^\top(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) + k\boldsymbol{\beta}^\top\boldsymbol{\beta},$$

قابل‌محاسبه است که در آن k پارامتر ستیغی نام دارد. برای محاسبه این پارامتر از روش اعتبارسنجی مقابل استفاده می‌شود. برآوردگر به روش ستیغی با محدودیت تصادفی برابر $\hat{\boldsymbol{\beta}}^{SR}_{Ridge} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + k\mathbf{I})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}}$ است.

[15] Cross-validation (CV)
[16] Leave-one-out cross-validation (LOCV)
[17] Stochastic linear restricted

۳ روش کمترین توان‌های دوم پیراسته

یکی از مشهورترین روش‌های موجود برای برآورد ضرایب رگرسیونی در داده‌هایی که دارای نقاط دورافتاده هستند روش کمترین توان‌های دوم پیراسته[۱۸] [۱۸] است که در ادامه تعریف می‌شود. برای محاسبه برآوردگر روش کمترین توان‌های دوم پیراسته کافی است مسئله

$$\min_{\boldsymbol{\beta},\mathbf{Z}} \varphi(\boldsymbol{\beta},\mathbf{Z}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top \mathbf{Z}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \quad s.t. \quad \mathbf{e}^\top\mathbf{z}=h, \quad z_i\in\{0,1\}, \quad i=1,...,n,$$

حل شود. که در آن $\mathbf{e}=(1,...,1)^\top_{n\times 1}$ و $\mathbf{z}=(z_1,\ldots,z_n)$ است. همچنین پارامتر h، پارامتر پیراسته[۱۹] است. لازم به ذکر است که ماتریس قطری $\mathbf{Z}$ با عناصر قطری z_i به‌منظور تعیین مشاهده دورافتاده و از بین بردن تأثیر آن روی مدل، در مسئله بهینه‌سازی حضور دارد، به‌طوری‌که در آن به مشاهدات دورافتاده وزن صفر اختصاص می‌یابد. یعنی،

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

تعریف می‌شود. لازم به ذکر است که در صد شکست (میزان استواری و مقاومت روش ذکرشده در برابر نقاط دورافتاده) این روش ۵۰ درصد است. یعنی حتی اگر نیمی از داده‌ها دورافتاده باشند بازهم این روش در برابر نقاط دورافتاده، استوار است [۱۳].

۱.۳ مدل‌های رگرسیونی کمترین توان‌های دوم پیراسته محدودشده

در شرایطی که به‌طور هم‌زمان با هم خطی و نقاط دورافتاده در داده‌ها مواجه باشیم ناچار به ترکیب روش ستیغی با روش کمترین دوم پیراسته خواهیم بود. در این شرایط برآورد ضرایب رگرسیونی با حل

$$\min_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{Z}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+k\boldsymbol{\beta}^\top\boldsymbol{\beta},$$

به دست می‌آید که در آن $\mathbf{Z}$ ماتریس قطری با عناصر قطری $\mathbf{z}=(z_1,\ldots,z_n)^\top, z_i\in\{0,1\}$ است. در این روش $z_i=0$ یعنی مشاهده i ام یک نقطه دورافتاده است.

به همین ترتیب برای برآورد ضرایب رگرسیونی مدل محدودشده تصادفی (۵) با فرض وجود هم خطی و نقطه دورافتاده در مجموعه داده کافی است مسئله

$$\min_{\boldsymbol{\beta}}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^\top\mathbf{Z}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})+k\boldsymbol{\beta}^\top\boldsymbol{\beta}.$$

را حل کنیم که در آن $\mathbf{Z}$ ماتریس قطری با عناصر قطری $\mathbf{z}=(z_1,\ldots,z_n)^\top, z_i\in\{0,1\}$ است. بدین طریق برآورد ضرایب رگرسیونی مدل محدودشده تصادفی با فرض وجود هم خطی و نقاط دورافتاده در مجموعه داده‌ها از رابطه

$$\hat{\boldsymbol{\beta}}^{SR}_{LTSRidge}=(\tilde{\mathbf{X}}^\top\mathbf{Z}\tilde{\mathbf{X}}+k\mathbf{I})^{-1}\tilde{\mathbf{X}}^\top\mathbf{Z}\tilde{\mathbf{y}}$$

محاسبه می‌شود که به‌طور مشابه فرم غیرتصادفی نیز قابل بازنویسی است.

۴ نتایج عددی

در این بخش مطالعه عددی روی داده‌های واقعی انجام می‌شود. در هر دو مجموعه داده ضرایب رگرسیونی به چهار روش کمترین توان‌های دوم معمولی (OLS)، روش ستیغی (Ridge)، روش کمترین توان‌های دوم پیراسته (LTS) و روش کمترین توان‌های دوم پیراسته ستیغی (LTSridge) در سه حالت بدون محدودیت، محدودشده خطی غیر تصادفی و محدودشده خطی تصادفی برآورد می‌شوند. نتایج نهایی با معیار مجموع توان‌های دوم خطا تحت ارزیابی قرار می‌گیرند.

۱.۴ مجموعه داده‌های واقعی

در این بخش مجموعه داده واقعی مربوط به مصرف آب موردبررسی قرار می‌گیرد.

۱.۱.۴ مجموعه داده آب

شرکت‌های آب و فاضلاب کشور حاوی داده‌های بسیار با اهمیت از حیث و نگاه مشترکین آب می‌باشد. طبق آیین‌نامه عملیاتی وزارت نیرو برای تعریف اشتراک آب، هر شخص حقیقی یا حقوقی که انشعاب آب یا انشعاب‌های (آب و یا فاضلاب) مورد تقاضای وی، طبق مقررات برقرارشده باشد و در تعریف دیگر هر شخص حقیقی یا حقوقی که برای مصرف آب شرب و سایر خدمات دریافتی از شرکت برای آن قبض صادر می‌گردد، در لیست مشترکین آب در شرکت‌های آب و فاضلاب ثبت می‌گردد و از خدمات این شرکت‌ها بهره‌مند خواهد گردید. کنتورهای نصب‌شده در محل اشتراک آب به‌صورت دوره‌ای کنتورخوانی می‌شوند (حدود شش الی هفت بار در سال) و مصارف هر یک از مشترکین جهت محاسبه مبلغ آب‌بها در سامانه‌های آبفا ثبت می‌گردد. داده‌های مورداستفاده در این پژوهش برحسب استانی بوده و به تفکیک سال‌های

[18] Least trimmed squares(LTS)

[19] Trimmed parameter

۱۳۸۱ الی ۱۳۹۸ یعنی بازه ۱۹ ساله است، بدین معنی که هر یک از متغیرها برحسب سال موردنظر جمع‌آوری گردیده است. قابل‌ذکر است که این داده‌ها مربوط به شهرهای استان سمنان می‌باشد و مشترکین روستایی شامل این داده‌ها نیست. در ادامه این بخش به تفسیر هر یک از متغیرها پرداخته می‌شود. متغیر مصرف آب نشان‌دهنده میزان مصرف کلیه مشترکین آب در کلیه کاربری‌ها طی یک سال است که برحسب واحد مترمکعب در سال ثبت می‌شود. این مقدار عددی میزان دقیق مصرف آب هر اشتراک می‌باشد که توسط کنتورخوان صورت گرفته و در سیستم جهت صدور قبوض ثبت می‌گردد. متغیر تعداد اشتراک نشان‌دهنده تعداد مشترکین موجود در استان در کلیه کاربری‌ها در پایان سال مذکور است که برحسب واحد فقره می‌باشد. لازم به ذکر است که فقط اشتراک‌های آب تنها موردنظر می‌باشد. متغیر جمعیت تعداد نفرات ساکن در شهرهای استان می‌باشد که برحسب پیش‌بینی حاصله از سرشماری‌های مرکز آمار برحسب نفر حاصل‌شده است. ازآنجاکه در سال‌های ۱۳۸۵، ۱۳۹۰ و ۱۳۹۵ درکشور سرشماری صورت گرفته است، لذا در این سال‌ها اعداد جمعیتی اعداد سرشماری است و مابقی سال‌ها برآورد جمعیت می‌باشد. متغیر تولید آب به معنای آب تولیدی از منابع تأمین آب برای شهرهای استانی است که برحسب مترمکعب در سال می‌باشد. به‌طورکلی منابع تأمین آب به دو دسته زیرزمینی و سطحی تقسیم می‌شود که در استان سمنان اکثر منابع تأمین آب زیرزمینی است. متغیر آب‌بها، بهای خالص دریافتی از مشترکین برحسب آب مصرفی مشترکین می‌باشد که برحسب پلکان‌های تعرفه آب مصرفی لحاظ شده است. لازم به ذکر است این متغیر شامل هیچ‌یک از موارد دیگر ذکرشده در قبوض نمی‌باشد. به‌عنوان‌مثال قبوض آب شامل مواردی نظیر آبونمان، تبصره‌ها، ماده ۱۱ فاضلاب‌بها و ... است. مدل این مجموعه داده‌ها به‌صورت

$$\begin{aligned}(Wateruse)_i &= \beta_1(Waterproduce)_i + \beta_2(Money)_i \\ +\beta_3(Year)_i &+ \beta_4(Crowd)_i + \beta_5(Branch)_i + \epsilon_i, \\ & i = 1, \ldots, 19, \end{aligned} \quad (۶)$$

است که در آن متغیر وابسته میزان مصرف آب $(Wateruse)$ است و متغیرهای توضیحی عبارت‌اند از سال آبی $(Year)$، مبلغ آب‌بها $(Money)$، جمعیت $(Crowd)$، میزان تولید آب $(Waterproduce)$ و تعداد انشعاب‌ها $(Branch)$[۲۰]. در ادامه به بررسی وجود هم خطی و نقاط دورافتاده در مجموعه داده‌ها پرداخته می‌شود. در شکل ۲ نحوه پراکنش نقاط دورافتاده نمایش داده‌شده است. نقاط دورافتاده با علامت "+" مشخص‌شده‌اند. نمودار انواع نقاط در مجموعه داده‌ها نیز در شکل ۳ قابل‌مشاهده است. طبق این نمودار مشاهدات ۱۲، ۱۴، ۱۶، ۱۸ و ۱۹ نقاط دورافتاده (پرت و اهرمی بد) هستند. مشاهدات ۱۹ و ۱۸ که به‌عنوان نقاط پرت معرفی‌شده‌اند، مربوط به داده‌های سال‌های ۱۳۹۷ و ۱۳۹۸ می‌باشد که با توجه به ثبات تقریبی مصرف آب، افزایش چند درصدی تعرفه آب، میزان آب‌بها رشد به سزایی داشته است. همچنین داده‌های ۱۲، ۱۴، ۱۶ مربوط به سال‌های قبل از هدفمندی یارانه‌ها است که مبلغ آب‌بها با تعرفه‌های قدیمی است.

معیار عامل تورم واریانس[۲۱] برای متغیرهای توضیحی در جدول ۱ نشان داده‌شده است. عامل تورم واریانس تمام متغیرهای توضیحی به‌جز متغیر $waterproduc$ بزرگ‌تر از ۱۰ است. همچنین مقدار آماره کاپا[۲۲] برابر $۱/۰۱۷۱۲۴ \times ۱۰^{۱۲}$ می‌باشد. نمودار ۴ همبستگی میان متغیرهای مجموعه داده را نمایش می‌دهد. این موارد نشان‌دهنده وجود هم خطی بالا در مجموعه داده‌ها است.

بر اساس برآوردگر کمترین توان‌های دوم معمولی و اطلاعات و مشاهدات موجود، ماتریس محدودیت به‌صورت

$$\mathbf{R} = \begin{bmatrix} ۰ & ۱ & ۳ & ۵/۰ & ۱ \\ ۰ & ۴ & ۱ & -۶/۰ & ۲ \\ ۱ & ۰ & -۱ & -۰/۵ & -۱ \end{bmatrix}, \quad (۷)$$

در نظر گرفته شد و بردار $\mathbf{r}$ به‌صورت تقریبی از رابطه

$$\mathbf{r} \simeq \mathbf{R}\hat{\boldsymbol{\beta}}_{ols} = \begin{bmatrix} ۱۸۰۹/۸۰۰ \\ ۳۹۴۵/۴۶۰ \\ -۱۸۲۰/۰۵۹ \end{bmatrix},$$

به دست آمد. بنابراین در مسائل کاربردی و عملی در نظر گرفتن محدودیت تصادفی خطی به‌جای محدودیت دقیق واقع‌بینانه‌تر به نظر می‌آید. اکنون لازم است فرض $H_0 : \mathbf{R}\boldsymbol{\beta} \simeq \mathbf{r}$ برای مدل (۶) آزمون شود. آماره آزمون این فرضیه به‌صورت

$$\mathcal{X}^2 \simeq (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top (\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^\top)^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) = ۰/۰۰۴۳۵۲۰۱۵,$$

است که در آن

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \\ \hat{\sigma}^2 &= \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \hat{\boldsymbol{\Sigma}} &= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}. \end{aligned}$$

[۲۰] منبع این داده‌ها شرکت آب و فاضلاب شهری استان سمنان می‌باشد.

[21] Variance Inflation Factor

[22] Kappa

بنابراین فرض صفر پذیرفته می‌شود. لازم به ذکر است که محدودیت‌های در نظر گرفته‌شده در مثال واقعی صرفاً بر مبنای اطلاعات پیشین و با استفاده از برآوردگر کمترین توان‌های دوم معمولی به‌دست‌آمده و به دلیل مطمئن شدن در مورد آن، از آزمون فرض محدودیت خطی که در [۱۲] آمده نیز استفاده‌شده است، تا علاوه بر پیاده‌سازی برآوردگرهای پیشنهادی جنبه آموزشی نیز داشته باشد.

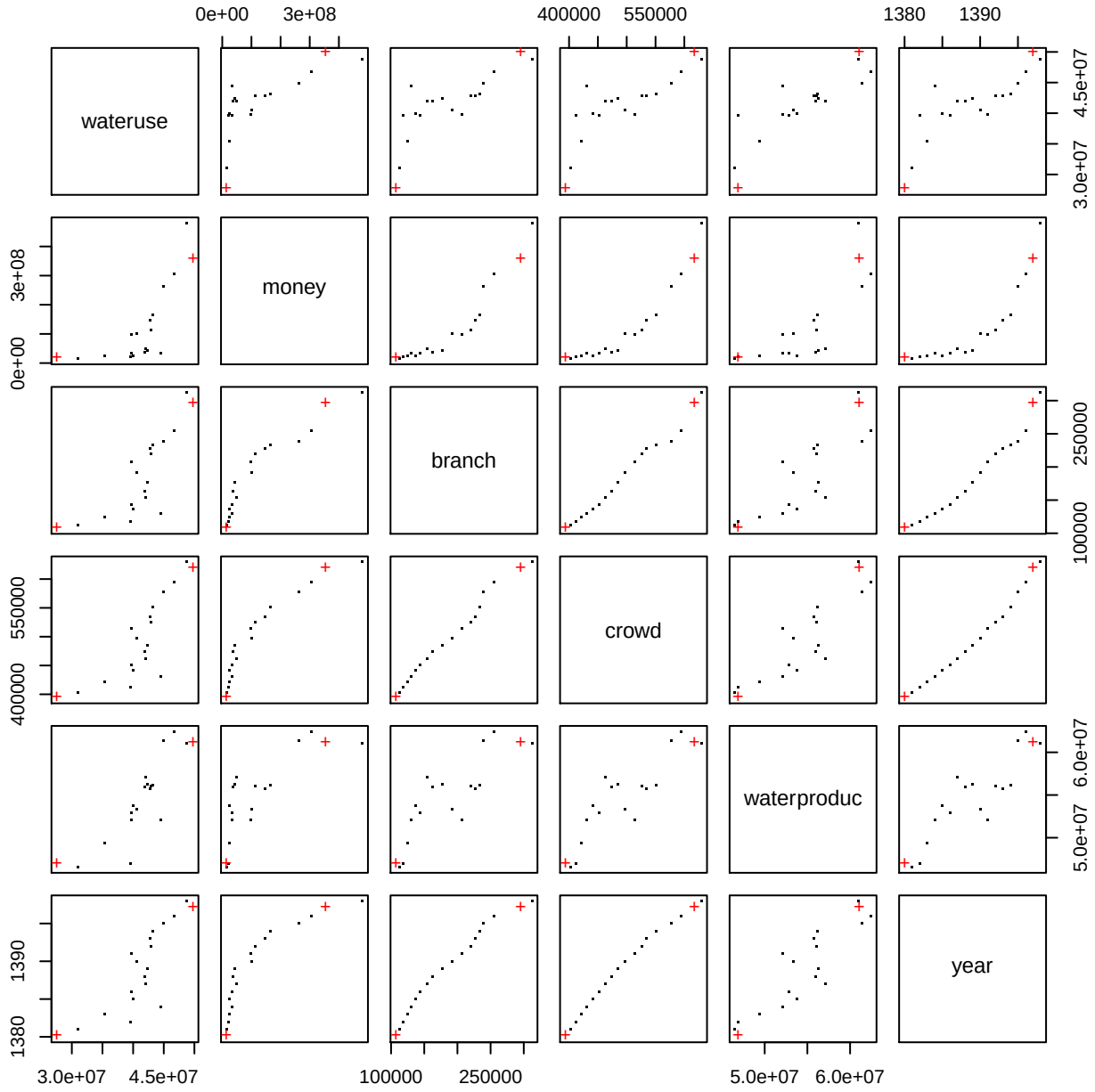
نتایج در جداول ۲ و ۳ گزارش‌شده‌اند. طبق جدول ۲ بهترین مدل برازش شده بر اساس معیار میانگین توان‌های دوم خطا، برآورد کمترین توان‌های دوم پیراسته ستیغی است. به‌عبارت‌دیگر برآورد رگرسیونی کمترین توان‌های دوم پیراسته ستیغی نقش پررنگ‌تری در توجیه واریانس کل دارد و میزان خطا در این مدل کمتر است که این امر کاملاً مورد انتظار بود چراکه در این مجموعه از داده‌ها به‌طور هم‌زمان نقاط دورافتاده و هم خطی وجود دارد. همچنین طبق جدول ۳ بهترین برآورد، برآورد کمترین توان‌های دوم پیراسته ستیغی محدودشده است که طبق معیار میانگین توان‌های دوم خطا نسبت به حالت غیر محدودشده، عملکرد بهتری دارد. بنابراین، بهترین برازش داده‌شده به متغیرهای آبی تعریف‌شده در این پژوهش که می‌تواند به‌عنوان برآورد پیش‌بینی مصرف آب به کار رود، برازش حاصل از روش محدودشده تصادفی کمترین توان‌های دوم پیراسته ستیغی است.

جدول ۱: عامل تورم واریانس برای متغیرها در مجموعه داده آب

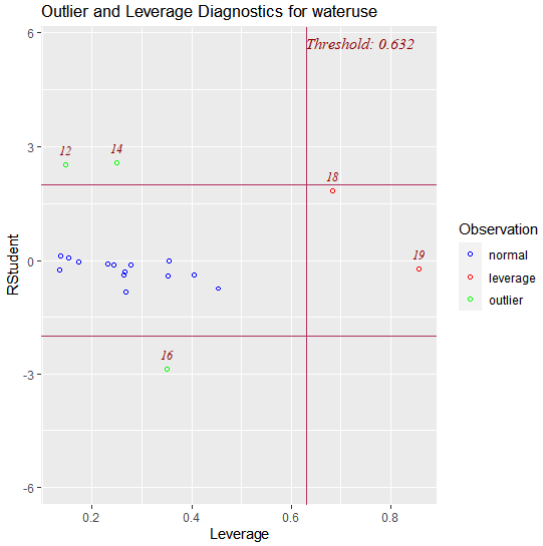
VIF_{year}	$\text{VIF}_{waterproduc}$	VIF_{crowd}	VIF_{branch}	VIF_{money}
۲۷۳٫۴۱۹۶	۹٫۶۰۶۰	۴۹۰٫۱۸۵۲	۱۳۱٫۲۷۲۴	۳۷٫۸۱۲۶

جدول ۲: برآوردگر ضرایب رگرسیونی برای مدل محدود نشده در مجموعه داده آب

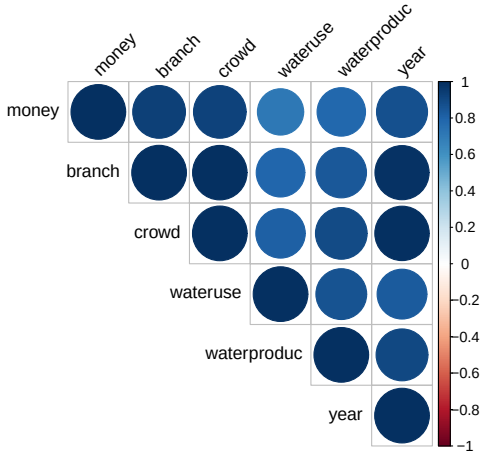
روش برآورد	OLS	Ridge	LTS	LTSRidge
$\hat{\beta}_1$	۰٫۳۹۲۷	−۰٫۳۸۵۸	−۰٫۰۰۶۴	−۰٫۴۵۳۷
$\hat{\beta}_2$	۰٫۵۰۱۶	−۰٫۵۱۲۹	۱٫۱۳۸۴	−۰٫۶۶۳۶
$\hat{\beta}_3$	−۲٫۱۶۰۸	۰٫۱۱۵۵	−۱٫۹۳۳۰۸	−۰٫۰۶۳۴
$\hat{\beta}_4$	۰٫۷۰۸۸	۰٫۲۱۳۳	۰٫۸۸۴۴	۰٫۷۱۹۲
$\hat{\beta}_5$	۱٫۴۹۷۲	۰٫۶۸۸۷	۰٫۹۹۳۰	۰٫۶۸۹۴
MSE	۱۵٫۵۳۰۱۲	۶٫۰۵۱	۱۰٫۹۹۷۷۳	۴٫۳۰۴
k	–	۰٫۰۰۴	–	۰٫۰۰۳



شکل ۲: نمودار پراکنش متغیرها در مجموعه داده مصرف آب



شکل ۳: نمودار انواع نقاط مجموعه داده مصرف آب



شکل ۴: نمودار همبستگی متغیرهای مجموعه داده آب

جدول ۳: برآوردگر ضرایب رگرسیونی برای مدل محدودشده خطی تصادفی در مجموعه داده آب

روش برآورد	OLS	Ridge	LTS	LTSRidge
$\hat{\beta}_1$	−۰٫۰۳۵۹	−۰٫۴۷۳۵	۰٫۰۰۳۵	−۰٫۴۶۸۳
$\hat{\beta}_2$	۰٫۳۸۵۳	−۰٫۲۱۷۵	۰٫۵۴۴۸	−۰٫۲۳۰۹
$\hat{\beta}_3$	−۰٫۴۹۸۳	۰٫۰۵۶۰	−۰٫۸۹۶۷	−۰٫۰۲۷۹
$\hat{\beta}_4$	۱٫۱۰۱۸	−۰٫۶۴۷۳	۱٫۰۰۱۲	−۰٫۵۱۳۵
$\hat{\beta}_5$	۰٫۰۵۹۲	۰٫۱۸۲۱	۰٫۳۹۳۳	۱٫۳۱۱۸
MSE	۲٫۹۵۱۰۷۹	۰٫۷۴	۰٫۳۱۶۱۸۱	۰٫۲۸۸
k	–	۰٫۰۰۲	–	۰٫۰۰۱

۵ بحث و نتیجه‌گیری

همان‌طورکه ذکر شد در مواردی‌که هم خطی و نقاط دورافتاده در مجموعه داده واقعی وجود داشته باشد، روش کمترین توان‌های دوم معمولی عملکرد مناسبی نخواهد داشت. روش ستیغی برای رفع مشکل هم خطی مناسب است اما زمانی که به همراه هم خطی نقاط دورافتاده در مدل وجود داشته باشند نیز روش مناسبی نخواهد بود و در این حالت لازم است که از روش کمترین توان‌های دوم پیراسته ستیغی که به‌طور هم‌زمان بر این دو مشکل غلبه می‌کند، استفاده کرد. طبق نتایج مجموعه داده‌های واقعی روش کمترین توان‌های دوم پیراسته ستیغی بهترین عملکرد را نسبت به سه روش دیگر داشته است. طبق نتایج داده‌های واقعی اعمال محدودیت روی مدل و به‌کارگیری اطلاعات غیر نمونه‌ای باعث بهبود نتایج و کاهش میانگین مجموع توان‌های دوم خطا شده است و این امر بدان معنی است که اعمال محدودیت باعث افزایش دقت مدل و نتایج خواهد شد.

تقدیر و تشکر

نویسندگان مقاله ضمن تشکر از اعضای محترم 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

[۱] امینی، م.، روزبه، م.، زمانی، ح. (۱۳۹۷)، **تحلیل رگرسیون پیشرفته با R**، انتشارات علمی پارسیان، تهران.

[۲] معنوی، م.، (۱۳۹۸)، *روش‌های تحلیل رگرسیونی در داده‌های با بعد بالا*، پایان‌نامه کارشناسی ارشد، دانشگاه سمنان، سمنان.

[3] Alheety, M. and Kibria, G. (2019). A new version of unbiased ridge regression estimator under the stochastic restricted linear regression model, *Communications in statistics: Simulation and computation*.

[4] Arashi, M., Kibria, B.M.G and Valizadeh T., (2017). On Ridge Parameter Estimators under Stochastic Subspace Hypothesis, *Journal Statistical computation and simulation*, **5**, 966-983.

[5] Belsley, D.A., Kuh, E., Welsch, R.E., (2004). *Regression Diagnostics Identifying Influential Data and Sources of Collinearity*, Wiley and Sons, New Jersey.

[6] Chatterjee, S. and Hadi, A. S., (1986). "Influential observation, High leverage points, and outliers in linear regression," *Journal Statistical Science*. **1** no.3, 379-416.

[7] Grob, J. (2003) Restricted ridge estimation, *Statistics & Probability Letters*, **65**, 57-64.

[8] Hald, A. (1952). *Statistical Theory with Engineering Applications*, NewYork: John Wiley.

[9] Hoerl A. E. and Kennard R. W. (1970), Ridge regression: biased estimation for non-orthogonal problems, *Thechnometrics*, **12**, 55-67.

[10] Kibria, B. M. G., (2005). Applications of some improved estimators in linear regression, *Journal Modern Applied Statistical Methods*, **5(2)**, 367-380.

[11] Liu H., Shah S. and Jiang W. (2004). On-line outlier detection and data cleaning, *Computers and Chemical Engineering*, **28 (9)**, 1635-1647.

[12] Montgomery, D.C., Peck, E. A. and Vining G. G. (2012) Introduction to Linear Regression Analysis, 5th Edition, John Wiley & Sons,New Jersey.

[13] Roozbeh M. (2016) Robust ridge estimator in restricted semi parametric regression models, *Journal of Multivariate Analysis*, **147**,127-144.

[14] Roozbeh M. and Aishah N. A. (2017). Feasible robust estimator in restricted semiparametric regression models based on the LTS approach, *Communications in Statistics – Simulation & Computation*, **46**, 7332-7350.

[15] Roozbeh M. and Aishah N. A. (2020). Uncertain stochastic ridge estimation in partially linear regression models with elliptically distributed errors, *Statistics: A Journal of Theoretical and Applied Statistics*, **54**, 494-523.

[16] Roozbeh, M. and Arashi, M. (2017). Least-trimmed squares: asymptotic normality of robust estimator in semiparametric regression models, *Journal of Statistical Computation & Simulation*. **147**, 1130-1147.

[17] 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*. **86(2)**, 357-372.

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

[19] Sarkar, N. (1992). A new estimator combining the ridge regression and the restricted least squares methods of estimation, *Communications in Statistics Theory Methods*, **21**, 1987-2000.

[20] Sengupta, D. and Jammalamadaka, S.R.(2003). *Linear Models: An Integrated Approach*. World Scientific Publishing Company.

[21] Theil, H., Goldberger, A.S., (1961). On pure and mixed statistical estimation in economics, *International Economic Review*, **2**, 65-78.

Application of stochastic restricted least trimmed squares ridge regression in water consumption modeling

Mahdi Roozbeh[1], Mlihe Malekjafarian[2] and Monireh Manavi[3]

Abstract:

The most important goal of statistical science is the analysis of the real data of the world around us. If this information is analyzed accurately and correctly, the obtained results will help us in many important decisions. Among the real data around us which its analysis is very important, is the water consumption data. Considering that Iran is located in a semi-arid climate area of the earth, it is necessary to take big steps for predicting and selecting the best and the most appropriate accurate models of water consumption, which is necessary for the macro-national decisions. In the analysis of the real data set, we usually encounter with the problems of multicollinearity and outliers points. Robust methods are used for analyzing the data sets with outliers and ridge method is used for analyzing the data sets with multicollinearity. Also, the restriction on the models is resulted from using non-sample information in estimation of regression coefficients. In this paper, it is proceeded to model the water consumption data using robust stochastic restricted ridge approach.

Keywords: Multicollinearity, Outliers, Ridge least trimmed squares method, Stochastic linear restriction, Water consumption.

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

[2]**Facing the Governor of Semnan Province City Water and Wastewater Company, Semnan, Iran.**

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