سیزدهمین کنفرانس بین المللی انجمن ایرانی تحقیق در عملیات
۳ و ۴ اردیبهشت ۱۳۹۹
13th International Conference of Iranian Operations Research Society
22-23 April 2020

بهینه‌سازی خطی ریج محدودشده تصادفی

منیره معنوی [1]، مهدی روزبه [2]

[1] کارشناس ارشد آمار ریاضي، دانشگاه سمنان دانشکده علوم ریاضی، آمار و علوم کامپیوتر؛
monireh.maanavi@semnan.ac.ir
[2] دانشیار دانشگاه سمنان، دانشگاه سمنان دانشکده علوم ریاضی، آمار و علوم کامپیوتر؛
mahdi.roozbeh@semnan.ac.ir

چکیده

در مدل رگرسیوني خطی اولین گام برای هرگونه تجزیه و تحلیلی، برآورد ضرایب رگرسیونی است. روش های گوناگونی از گذشته تا کنون نظیر کمترین قدر مطلق انحرافات معرفی شد. اما ساده ترین و کاربردی ترین روش برای برآورد ضرایب روش کمترین توان های دوم معمولی است که تحت قضیه گاوس-مارکف این برآوردگر، بهترین برآوردگر خطی نااریب با کمترین واریانس است. متاسفانه روش کمترین توان های دوم معمولی در صورت وجود ارتباط خطی بین متغیرهای توضیحی روش مطمئنی نخواهد بود. بدین سبب مستلزم بکارگیری روش های دیگری هستیم. روش ریج یکی از محبوب‌ترین روش‌ها برای رفع مشکل همخطی است. در واقع روش ریج یک تبدیل خطی از روش کمترین توان‌های دوم می‌باشد که خطای کمتری را نسبت به آن روش دارد.

در بعضی شرایط بین عناصر بردار پارامترها روابطی برقرار است که در صورت درنظر نگرفتن آن‌ها در درستی برآورد به‌دست آمده تردید به وجود می‌آید. هدف اصلی در این مقاله ارائه یک برآورد در مدل‌های رگرسیون خطی محدود شده تصادفی با خطای ناهمبسته، زمانی که هم‌خطی چندگانه در ماتریس طرح وجود دارد، است. به همین سبب ابتدا برآوردگرهای محدودشده و سپس برآورد ریج را درمدل‌های رگرسیون خطی معرفی نموده و برآورد را براساس روش ریج تحت محدودیت تصادفی به‌دست می‌آوریم. در انتها، عملکرد برآوردگر پیشنهادی را با استفاده از مجموعه داده شبیه‌سازی‌شده مورد بررسی قرار می‌دهیم.

کلمات کلیدی: برآورگر ریج پیراسته؛ خطای همبسته؛ کمترین توان‌های دوم پیراسته؛ مدل رگرسیونی محدودشده تصادفی؛ هم‌خطی.

مقدمه

مدل رگرسیون خطی به‌صورت

$$y=X\beta+\varepsilon(1)$$

است که در آن $y=(y_1,\dots,y_n)^T$ بردار مشاهدات متغیر پاسخ، $X=(x_1,\dots,x_p)$ ماتریس مشاهدات متغیرهای توضیحی با $x_i=(x_{1i},\dots,x_{ni})^T$ به ازای $i=1,\dots,p$، بردار ضرایب رگرسیونی و $\beta=(\beta_1,\dots,\beta_p)^T$ بردار خطای تصادفی با شرایط $E(\varepsilon)=0$ و $E(\varepsilon\,\varepsilon^T)=\sigma^2 I_p$ است. به منظور برآورد ضرایب رگرسیونی به روش کمترین توان های دوم معمولی تابع مربع خطا به‌صورت $(y-X\beta)^T(y-X\beta)$ نسبت به β کمینه می شود و برآوردگر برابر $\hat{\beta}_{OLS}=(X^TX)^{-1}X^Ty$ است.

در عین سادگی محاسبه و کاربرد فراوان، این روش در شرایطی نظیر هم‌خطی[1] در داده‌ها عملکرد ضعیفی داشته و در درستی جواب‌های آن ابهاماتی به وجود می‌آید. در چنین شرایطی به‌کارگیری روش‌های دیگری نظیر رگرسیون ریج و درصورت وجود نقاط دورافتاده رگرسیون استوار پیشنهاد می‌شود. هورل و کنارد[1] اولین بار از رگرسیون ریج برای رفع مشکل همخطی استفاده نمودند. رگرسیون ریج، یکی از روش‌های جریمه‌ای است که در آن تابع هدف روش کمترین توان‌های دوم جریمه می‌شود و همین امر باعث اصلاح جواب‌های روش کمترین توان‌های دوم می‌شود.

وجود همخطی ممکن است منجر به فواصل اطمینان وسیع برای پارامترها ، ناپایداری پارامترها

Collinearity 1

و یا تولید برآوردهایی با علامت اشتباه شود. روش‌های مختلفی برای غلبه برای همخطی وجود دارد از جمله رگرسیون مولفه‌ی اصلی ، رگرسیون ریج و ... .
یکی دیگر از نکات بسیار مهم در برآورد پارامترها یافت روابط موجود میان متغیرهای توضیحی و بکارگیری آن‌ها در برآورد است. زیرا این روابط روی مقادیر به دست آمده تاثیر زیادی خواهند داشت. بنابراین نادیده گرفتن این روابط، درستی برآوردهای به دست آمده را تحت الشعاع قرار می‌دهد. بدین جهت می‌بایست پارامترهای مدل، تحت این روابط محدود شوند. مدل‌های محدود شده به طور گسترده‌ای در مسئله آزمون فرضیه، به ویژه آزمون نسبت درست نمایی تعمیم یافته[2] در مدل‌های رگرسیون کاربرد دارند. البته مدل‌های محدودشده خود، به دو گروه تصادفی و غیر تصادفی تقسیم می‌شوند. به بیان دیگر می‌توان گفت: استفاده از اطلاعات غیر نمونه‌ای در مدل‌های آماری معمولا به‌صورت یک محدودیت مطرح می‌شود. محدودیت می‌تواند به یکی از دلایل
(1) حقیقت ناشی از ملاحظات نظری و یا تجربی (آزمایشگاهی)، (2) فرضیه‌ای که باید آزمون شود، (3) نظر محقق یا متخصص، (4) شرط تحمیل‌شده غیرمعمول برای کاهش یا حذف فزونگی در مدل مورد بررسی به وجود آید[2]. با اعمال محدودیت‌ها در حقیقت از اطلاعات اضافی نمونه و حتی در برخی موارد نمونه‌های قبلی استفاده می‌شود و پر واضح است که هر چه از اطلاعات نمونه‌ای بیشتر استفاده شود دقت برآوردگرها افزایش می‌یابد. این محدودیت ممکن است به‌صورت غیر تصادفی، یعنی $r=R\beta$ بیان شود، به‌طوری که r بردار q بعدی با عناصر معلوم و R ماتریس $q\times p$ اطلاعات پیشین معلوم روی بردار پارامترها است. ماتریس R سطری رتبه کامل است. اما حقیقت این است که در عمل، موقعیت‌های بسیاری وجود دارد که در آن فرض می‌شود که اطلاعات پیشین مناسب نیستند یا به بیان ساده‌تر، در عمل اطلاعات پیشین دقیق نیستند[3]. محدودیت خطی تصادفی در مواردی نظیر روابط اقتصادی، ساختارهای صنعتی، برنامه‌ریزی‌های تولیدی و ... کاربرد دارند. اگر عدم اطمینان در مورد اطلاعات پیشین وجود داشته باشد، یک جایگزین مناسب برای محدودیت‌های خطی، محدودیت‌های خطی تصادفی[3] به‌صورت $r=R\beta+v$ است به‌طوری که r بردار q بعدی با عناصر معلوم و R ماتریس $q\times p$ طلاعات پیشین معلوم روی بردار پارامترها و v خطای تصادفی (اطلاعات پیشین مجهول) است. تیل و گلدبرگر [4] با ترکیب محدودیت خطی تصادفی و مدل (1) مدل جدید با فرض استقلال v و ε به‌صورت

$$\begin{bmatrix} y \\ r \end{bmatrix}=\begin{bmatrix} X \\ R \end{bmatrix}\beta+\begin{bmatrix} \varepsilon \\ v \end{bmatrix}(2)$$

به‌دست آوردند که می‌توان به‌صورت

$$\tilde{y}=\tilde{X}\beta+\tilde{\varepsilon}(3)$$

بازنویسی کرد که در واقع، همان مدل رگرسیونی خطی ساده با شرایط $E(\tilde{\varepsilon})=0$ و $E(\tilde{\varepsilon}\tilde{\varepsilon}^T)=\sigma^2 I_{(n+q)\times p}$ است. که ضرایب آن به سادگی با روش کمترین توان‌های دوم معمولی محاسبه است و آن را برآوردگر محدودشده تصادفی کمترین توان‌های دوم $\hat{\beta}_{OLS}^{SR}=(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{y}$ می‌نامیم.
هدف اصلی در این مقاله ارائه یک برآورد ریج در مدل رگرسیونی خطی محدود شده تصادفی است. در بخش 1، برآوردگر ریج با محدودیت خطی تصادفی، در بخش 2، کاربرد برآوردگر معرفی شده در داده‌های سیمان مطرح شده است.

1- برآوردگر ریج با محدودیت خطی تصادفی

هورل و کنارد *[1]* از نخستین افرادی بودند که با بکارگیری روش ریج برای رویارویی با این مشکل مبادرت ورزیدند. پس از تلاش‌های آن دو این روش بیش از پیش رواج یافت. طبیعتا با حضور همخطی در مدل بین ستون‌های ماتریس $X^T X$ ارتباط خطی وجود دارد و در این شرایط طبق خواص دترمینان، دترمینان این ماتریس بسیار کوچک شده و به سمت صفر میل می‌کند، بنابراین درایه‌های وارون ماتریس $X^T X$ بسیار بزرگ شده و دقت برآوردگر کمترین توان‌های دوم بسیار پایین می‌آید و قابل استناد نخواهد بود. بزرگترین مزیت روش ریج وارون‌پذیری

GLRT 2
Stochastic linear constraints 3

همیشگی ماتریس $X^T X$ است. مسئله بهینه سازی روش ریج با فرض ناهمبسته بودن خطا به صورت

$$\min_{\beta}(y-X\beta)^T(y-X\beta)+k\beta^T\beta$$

است که در آن k پارامتر ریج نام دارد. برای محاسبه این پارامتر از روش هایی مختلفی می توان استفاده کرد که در این مقاله از روش اعتبارسنجی مقابل استفاده می شود. برآوردگر به روش ریج برابر $\hat{\beta}_{Ridge}=(X^T X+k I_p)^{-1}X^T y$ است.

مسئله کمینه‌سازی ریج با فرض ناهمبسته بودن خطاها و در نظر گرفتن محدودیت تصادفی به‌صورت زیر است.

$$\min_{\beta}(y-X\beta)^T(y-X\beta)+k\beta^T\beta$$

$$r=R\beta+v$$

برآوردگر به روش ریج محدودشده برابر $\hat{\beta}^{SR}_{Ridge}=(\tilde{X}^T\tilde{X}+k I_p)^{-1}\tilde{X}^T\tilde{y}$ است.

مقالات *[6]* و *[7]* منابع بسیار مفیدی در این خصوص هستند که به خواننده علاقه‌مند پیشنهاد می‌شود.

2- مطالعه شبیه سازی

در این مطالعه 100 مشاهده از مدل رگرسیونی $y=X\beta+\varepsilon$ با در نظر گرفتن β به صورت $\beta=(-2,3,5,-6,4)$ و $\varepsilon \sim N(0,1)$ است. به منظور ایجاد همخطی بین متغیرهای توضیحی درایه‌های ماتریس طرح به‌صورت

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

شبیه‌سازی شده است. به‌طوری که در آن z_{ij} اعداد شبه تصادفی نرمال استاندارد مستقل بوده و γ طوری تعیین می‌شود که بین هر دو متغیر توضیحی همبستگی ایجاد شود. در این مطالعه به‌ازای دو مقدار $\gamma=0.78$ و $\gamma=0.9$ درایه‌ها تولید می‌شوند. لازم به ذکر است که به منظور ایجاد محدودیت خطی تصادفی موارد

$$\begin{bmatrix}1&4&7&10&13\\2&5&8&11&-2\\3&6&9&12&15\end{bmatrix},\eta \sim X^2(10),r=R\beta+\eta,r=\begin{bmatrix}45.9784\\-10.9318\\54.2189\end{bmatrix}$$

در نظر گرفته شد. لازم به ذکر است که تعداد تکرار حلقه‌های شبیه‌سازی مونت کارلو 1000 است. جدول برآورد پارامترها به روش کمترین توان‌های دوم محدودشده و روش ریج محدوشده به ازای دو مقدار 0.9و $\gamma=0.87$ در در جدول 1 گزارش شده است. پرواضح است که روش ریج در هر دو مورد عملکرد بهتری داشته است.

نمودارهای همبستگی میان متغیرهای توضیحی به ازای $\gamma=0.87$ در شکل 1 رسم شده است. این نمودارها با استفاده از نرم‌افزار R رسم شده است. وجود همخطی به وضوح در نمودارها دیده می‌شود. بیشترین میزان همبستگی برابر 89/0 است که مابین متغیرهای x_1 و x_5 ، متغیرهای x_2 و x_5 است.

نمودارهای همبستگی میان متغیرهای توضیحی به ازای $\gamma=0.87$ در شکل 2 رسم شده است. وجود همخطی به وضوح در نمودارها دیده می‌شود. بیشترین میزان همبستگی برابر 94/0 است که مابین متغیرهای x_4 و x_5 است.

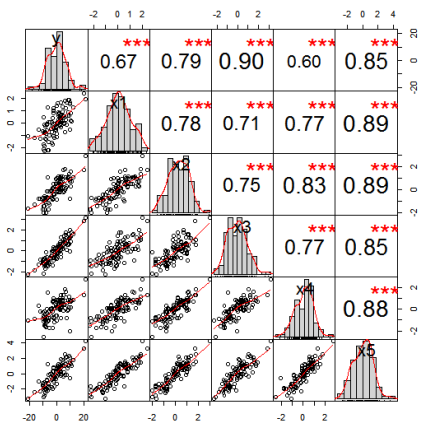
جدول 1: ارزیابی برآوردگرها به‌ازای $\gamma=0.87$ و $\gamma=0.9$

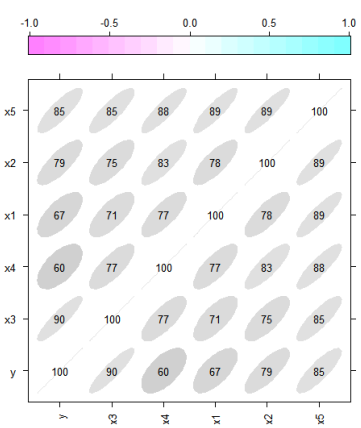
$\gamma=0.87$

روش	کمترین توان‌های دوم معمولی محدودشده			ریج محدودشده		
ضریب رگرسیونی	برآورد	اریبی	انحراف استاندارد	برآورد	اریبی	انحراف استاندارد
$\hat{\beta}_1$	2558/-2	2558/-0	2571/0	-9045/1	0945/0	1968/0
$\hat{\beta}_2$	8611/2	1388/-0	3041/0	9880/2	-0119/0	1084/0
$\hat{\beta}_3$	2186/5	2186/0	1123/0	9067/4	-0923/0	2261/0
$\hat{\beta}_4$	3737/-5	6262/0	1744/0	-8178/5	1821/0	1197/0
$\hat{\beta}_5$	9475/3	0524/-0	2935/0	8743/3	1256/0	2053/0
میانگین مربعات خطا	4262/1			9235/0		

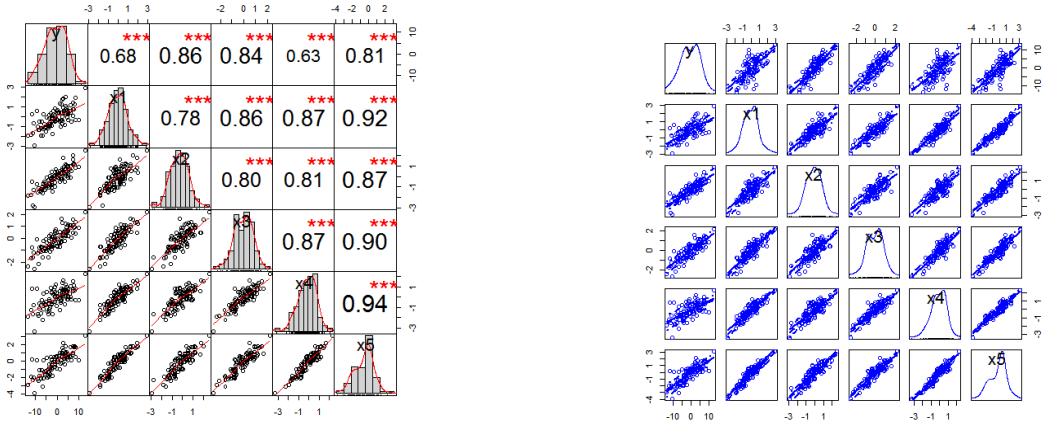
$\gamma=0.9$

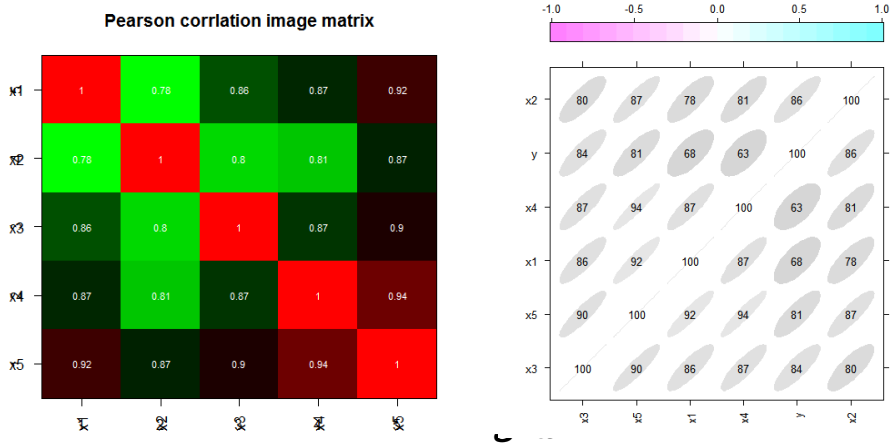
روش	کمترین توان‌های دوم معمولی محدودشده			ریج محدودشده		
ضریب رگرسیونی	برآورد	اریبی	انحراف استاندارد	برآورد	اریبی	انحراف استاندارد
$\hat{\beta}_1$	6933/-2	6933/-0	2787/0	-8226/1	1773/0	2781/0
$\hat{\beta}_2$	8119/2	1880/-0	2684/0	0414/3	0414/0	2333/0
$\hat{\beta}_3$	3060/5	3060/0	2617/0	9841/4	-0158/0	2460/0
$\hat{\beta}_4$	1898/-5	8101/0	2713/0	-6813/5	3186/0	3238/0
$\hat{\beta}_5$	9987/3	0012/-0	3036/0	6270/3	-3729/0	3829/0
میانگین مربعات خطا	5772/1			0013/0		





شکل 1: نمودارهای بررسی ارتباط خطی میان متغیرهای توضیحی به‌ازای $\gamma = 0.87$





2: نمودارهای بررسی ارتباط خطی میان متغیرهای توضیحی به‌ازای $\gamma = 0.9$

3- نتیجه و جمع‌بندی

در این مقاله مسئله بهینه‌سازی کمترین توان‌های دوم معمولی محدودشده و ریج محدودشده در مجموعه داده شبیه‌سازی شده با حضور هم‌خطی حل شد. نتایج بر عملکرد بهتر ریج محدودشده گواهی می‌دهند.

4- مراجع

[1] Hoerl A. E. and Kennard R. W. ; "*Ridge regression:Some simulatione*", Comm. Statist. Simulation Comput.4, 105–123, (1975).

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

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

[4] Arashi, M., Golam Kibria, B.M. and Valizadeh T.; *On ridge parameter estimators under stochastic subspace hypothesis*, Journal Statistical computation and simulation, **5**, 966–983, (2017).

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

[6] Fallah, R., Arashi, M. and Tabatabaey, S. M. M. ; "*On the ridge regression estimator with sub-space restriction*", Communication in Statistics Theory and Methods, **46**, 11854-11865, (2017).

[7] Karbalaee, M. H., Arashi, M. and Tabatabaey, S. M. M.; "*Performance analysis of the preliminary test estimator with series of stochastic restrictions*", Communication in Statistics Theory and Methods, **47**, 1-17, (2018).