Kharazmi University

Mathematical Researches

Print ISSN: 2588-2546
Online ISSN: 2588-2554

Homepage: https://mmr.khu.ac.ir/

A Simulation Study for Comparing Ridge and Lars Regression Estimators

Roshanak Alimohammadi[1], Zhaleh Bahari[2]

1. Faculty of Mathematical Sciences, Department of Statistics, Alzahra University, Tehran, Iran.

✉ E-mail: r_alimohammadi@alzahra.ac.ir

2. Faculty of Mathematical Sciences, Department of Statistics, Alzahra University, Tehran, Iran.

E-mail: jaleh.bahari90@yahoo.com

Article Info

Article type:

Research Article

Article history:

Received:
7 May 2019
Revised form:
18 January 2020
Accepted:
14 December 2020
Published online:
21 May 2022

Keywords:

Ordinary Least Squares Error Regression;
Ridge Ression;
Bridge Regression;
Lasso Regression;
LARS Regression;
Tuning Parameter.

ABSTRACT

Introduction

Regression analysis is a common method for modeling relationships between variables. Usually Ordinary Least Squares method is applied to estimate regression model parameters. These estimators are unbiased and appropriate when design matrix is nonsingular.

In presence of multicollinearity, design matrix is singular and Ordinary Least Squares estimates cannot be obtained. In this situation, other methods, such as Lasso, Ridge and Lars may be considered.

Other hand, in many fields such as medicine, number of variables is greater than the number of observations and usual methods such as Ordinary Least Squares are not proper and shrinkage methods, such as Lasso, Ridge and ... have a better performance to estimate regression model coefficients. In the shrinkage methods, tuning parameter plays an essential role in selecting variables and estimating parameters. Bridge shrinkage estimators is an estimator that can be obtained by changing its tuning parameter. In other words, Bridge method is the extension of Ridge and Lasso regression methods. Selecting the appropriate amount of tuning parameter is important. There are many studies on each of these methods under the assumed conditions. In this paper, performance of Bridge shrinkage estimators, such as Lasso and Ridge are compared with Lars and Ordinary Least Squares estimators in a simulation study.

Material and Methods

A simulation study is applied to compare performance of the regression methods Ridge, Lasso, Lars and Ordinary Least Squares. MSE criterion is applied to evaluate the method performance.

Statistical software R is applied for simulation and comparing the regression methods.

Results and discussion

In the presence of collinearity, Bridge regression estimators will result in appropriate estimators. These estimators are biased but

their performance is better than unbiased estimators such as Ordinary Least Squares. Indeed, Bridge estimators have the best performance in the class of biased estimators.

Conclusion

In this article, Ridge and Lasso estimators as special cases of Bridge estimators are compared with Lasso and Ordinary Least Squares in a simulation study.

This study shows that under the supposed conditions, Ridge, Lasso and Lars have better action than Ordinary Least Squares method.

Lars method has the best performance and Ridge estimators is better than Lasso Regression.

How to cite: Alimohammadi, R., Bahari, Zh., (2022) A Simulation Study for Comparing Ridge and Lars Regression Estimators. *Mathematical Researches*, 8 (2), 1-13

 Publisher: Kharazmi University

شاپا چاپی: ۲۵۸۸-۲۵٤٦
شاپا الکترونیکی: ۲۵۸۸-۲۵٥٤

پژوهش‌های ریاضی

Homepage: https://mmr.khu.ac.ir/

مطالعهٔ شبیه‌سازی برای مقایسه برآوردگرهای رگرسیونی بریج و لارس

روشنک علیمحمدی[۱]✉، ژاله بهاری[۲]

۱. نویسندهٔ مسئول، گروه آمار، دانشکدهٔ ریاضی، دانشگاه الزهرا، تهران، ایران. پست الکترونیکی: r_alimohammadi@alzahra.ac.ir

۲. گروه آمار، دانشکدهٔ ریاضی، دانشگاه الزهرا، تهران، ایران. پست الکترونیکی: jaleh.bahari90@yahoo.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۳۹۸/۰۲/۱۷
تاریخ بازنگری: ۱۳۹۸/۱۰/۲۸
تاریخ پذیرش: ۱۳۹۹/۰۸/۲۴
تاریخ انتشار: ۱۴۰۱/۰۲/۳۱

واژه‌های کلیدی:
ترتیب هسیان،
مخروط محدب،
ترتیب محدب خطی،
توزیع هذلولوی تعمیم‌یافته.

چکیده

تحلیل رگرسیون یکی از روش‌های متداول آماری در مدل‌سازی روابط بین متغیرهاست. لذا در رگرسیون دو موضوع تعیین روابط بین متغیرها و تحلیل روابط حاصل مورد توجه قرار می‌گیرد.

در مسائل با بعد بالا وقتی تعداد متغیرها بیشتر از تعداد مشاهدات است، روش‌های معمول مانند رگرسیون کمترین توان‌های دوم عادی کارایی لازم را ندارند و روش‌های انقباضی، ازجمله لاسو، ریج و ... از کارایی بهتری در برآورد ضرایب رگرسیونی برخوردار هستند. در این برآوردگرها پارامتر کنترل نقش اساسی در انتخاب متغیرهای تبیینی و برآورد ضرایب مدل بازی می‌کند. برآوردگرهای انقباضی بریج، برآوردگری است که با تغییر پارامتر کنترل آن می‌توان به برآوردگرهای ریج و لاسو دست یافت. در این مقاله برآوردگر انقباضی بریج از جمله لاسو و ریج را با برآوردگر لارس و کمترین توان‌های دوم معمولی مقایسه کرده و کارایی آنها را با معیار میانگین توانهای دوم خطا مورد ارزیابی قرار می‌دهیم.

استناد: علیمحمدی، روشنک؛ بهاری، ژاله؛ (۱۴۰۱). مطالعهٔ شبیه‌سازی برای مقایسه برآوردگرهای رگرسیونی بریج و لارس. *پژوهش‌های ریاضی*، ۸ (۲)، ۱-۱۳.

ناشر: دانشگاه خوارزمی

۱. مقدمه

تحلیل رگرسیونی تکنیکی آماری برای بررسی و به مدل درآوردن ارتباط بین متغیرهاست. در بسیاری از مسائل از جمله پزشکی تعداد متغیرهای تبیینی (p) بسیار زیاد و تعداد مشاهدات (n) کم است که به مساله p بزرگ و n کوچک معروف است. این مساله باعث بروز اشکال در برازش مدل‌های رگرسیون می‌شود. از طرفی برخی از متغیرها رفتار نسبتا یکسانی با متغیرهای دیگر دارند یا در واقع برخی از متغیرها ترکیب خطی از یک یا چند متغیر دیگر هستند. از این رو در رویارویی با این گونه مسائل، یک زیرمجموعه کوچک از متغیرها که دارای بیش‌ترین تأثیر بوده را انتخاب کرده و به برآورد ضرایب آنها می‌پردازیم.

انتخاب متغیر و برآورد کردن ضرایب آنها اساسی‌ترین بخش در مدل‌سازی رگرسیونی است. روش‌های برآوردیابی رگرسیون کمترین توان‌های دوم عادی[1] (OLS)، انتخاب متغیرهای تبیینی به‌صورت گام به گام و ... در مواجهه با داده‌هایی که از ویژگی‌های متفاوتی برخوردار باشند عملکرد قابل اطمینانی از خود نشان نمی‌دهند. از آسیب‌های مدل در هنگام استفاده از این روش‌ها می‌توان به عدم پایداری، دقت پیش‌بینی کم و انتخاب نادرست متغیرها اشاره نمود. به‌علاوه این مشکلات زمانی که همبستگی بین متغیرهای تبیینی زیاد باشد تشدید نیز می‌شوند. روش‌های انقباضی به‌عنوان راهکاری در جهت کاهش این مشکلات به‌خصوص وقتی همبستگی بین متغیرهای تبیینی زیاد باشد مورد توجه قرار گرفته‌اند.

جیمز و استاین[2] [۵] نشان دادند که روش‌های انقباضی برای تابع زیان توآنهای دوم خطا، کارایی برخی برآوردگرها را افزایش می‌دهند. در این روش‌ها ضرایب رگرسیونی را با اعمال محدودیت روی دامنه تغییرات آنها برآورد می‌کنند. اگرچه وجود چنین محدودیت‌هایی واریانس برآوردگر را کاهش می‌دهد ولی مقداری اریبی ایجاد می‌کند، به طوری‌که می‌توان امیدوار بود در نهایت میانگین توان‌های دوم خطا (MSE) کاهش یابد.

۲. مروری بر روش‌های برآوردیابی ضرایب مدل رگرسیون

مدل رگرسیون خطی به صورت ماتریسی

$$Y = X\beta + \epsilon$$

را در نظر بگیرید، که در آن $Y_{(n\times 1)}$ بردار مشاهدات متغیر پاسخ، $X_{(n\times p)}$ ماتریس مقادیر متغیرهای تبیینی، $\beta_{(p\times 1)}$ بردار ضرایب رگرسیونی و $\epsilon_{(n\times 1)} = (\epsilon_1, \dots, \epsilon_n)^T$ بردار خطای تصادفی با امید $E(\epsilon) = 0$ و کوواریانس $Cov(\epsilon) = \sigma^2 \mathrm{I}$ و $(n, p) \geq 1$ است. در این مدل jامین ستون ماتریس X با x_j و i امین عنصر x_j با x_{ij} نمایش داده می‌شود. برآوردگر کمترین توان‌های دوم معمولی برآوردگری است که مجموع توان‌های دوم مانده‌ها یعنی $Q = Q(\beta_1, \dots, \beta_p) = \epsilon^T\epsilon = \sum_{i=1}^{n} \epsilon_i^2$ را مینیمم نماید. در صورت ناتکین بودن ماتریس X^TX ، با مشتق‌گیری از Q

[1] Ordinary Least Squares Estimator

[2] James and Stein

نسبت به β و برابر صفر قرار دادن مشتق حاصل، برآوردگر ضرایب مدل به‌صورت $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$ حاصل می‌شود.

۱،۲ روش انقباضی بریج

فرانک و فریدمن[1] [۳] مجموع توان‌های دوم ماندهٔ Q را تحت قید $\sum_{j=1}^{p} |\beta_j|^{\gamma} \leq t$ ، به ازای $t \geq 0$ و $\gamma \geq 0$ مینیمم کردند و آن را به‌عنوان تعمیمی از رگرسیون ریج[2] و لاسو[3] به نام رگرسیون بریج[4]، برای مقابله با همخطی چندگانه ارائه نمودند. برآوردگر حاصل از حل مساله بهینه‌سازی فوق به صورت

$$\hat{\beta}_{Bridge} = \underset{\beta}{\underbrace{argmin}} \; ((Y - X\beta)^T (Y - X\beta) + \mathrm{k} \sum_{j=1}^{p} |\beta_j|^{\gamma})$$

حاصل می‌شود، که در آن k پارامتر کنترل[5] بوده و میزان انقباض تحمیل شده به ضرایب را کنترل می‌کند. لازم به ذکر است که انتخاب بهینه γ منجر به افزایش کارایی برآوردگر خواهد شد. در این رابطه اگر $\gamma = 2$ ، مساله فوق به رگرسیون ریج تبدیل می‌شود و اگر $\gamma = 1$، رگرسیون لاسو است. مالیک و یی[6] [۸] رگرسیون بریج بیزی را مورد مطالعه قرار دادند.

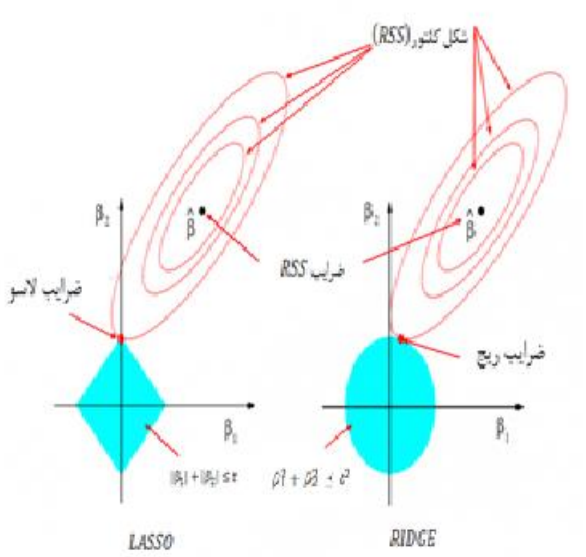
برای $\gamma = 2$ برآوردگر ریج دارای صورت بسته $\hat{\beta}_{Ridge} = (X^T X + kI)^{-1} X^T Y$ است. برخلاف روش ریج، لاسو صورت بسته‌ای ندارد و برای هر مقدار k بردار ضرایب به‌صورت عددی به‌دست می‌آید، اما تیبشیرانی[7] [۱] نشان داد اگر ماتریس طرح متعامد[8]، یعنی $X^T X = I$ باشد، ضرایب برآوردگر لاسو به صورت

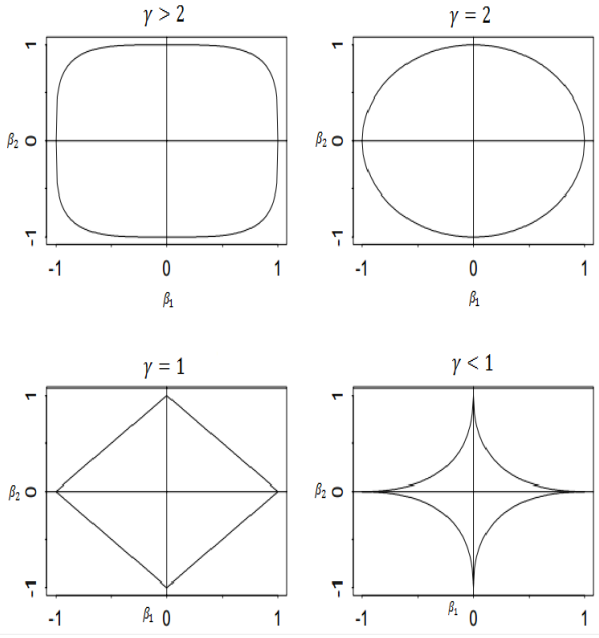
$$\hat{\beta}_{Lasso} = Sgn(\hat{\beta}_{OLS}) \left(\hat{\beta}_{OLS} - \frac{k}{2} I \right)^{+}$$

به دست می‌آید، که در آن Sgn تابع علامت و $a^{+} = \max(0, a)$ است.

برآوردگر بریج در حالت کلی همانند لاسو دارای صورت بسته‌ای نیست.

[1] Frank and Friedman
[2] Ridge
[3] LASSO (Least Absolute Shrinkage and Selection Operator)
[4] Bridge Regression
[5] Tuning Parameter
[6] Mallick and Yi
[7] Tibshirani
[8] Orthogonal





شکل۱. عملکرد رگرسیون بریج در حالت $p = 4$ و $t = 1$

شکل۱ ناحیهٔ تاوان $(|\beta_1|^\gamma + |\beta_2|^\gamma \leq 1)$ و مجموع توان دوم مانده‌ها با مرکزیت برآورد کمترین توان‌های دوم عادی را نشان می‌دهد. این شکل به وضوح ناتوانی روش بریج درحالت $\gamma \geq 2$ در صفر برآورد کردن ضرایب را نشان داده، که برخورد دو ناحیه نمی‌تواند در نقطه‌ای باشد که در آن یکی از ضرایب صفر است. اما وقتی $\gamma \leq 1$، برآورد ضرایب به گونه‌ای انجام می‌پذیرد که بعضی از ضرایب، مربوط به متغیرهای بی‌اثر دقیقاً صفر برآورد شده و به این ترتیب متغیر مربوط به آن ضرایب از مدل خارج خواهد شد. بنابراین در این روش، برآوردیابی و انتخاب متغیر تواماً صورت می‌پذیرد،

که این ویژگی استفاده از روش انقباضی بریج را برای $\gamma \leq 1$ (از جمله لاسو) در بعد بالا بسیار محبوب ساخته است. برای محاسبهٔ واریانس برآوردگر بریج می‌توان به فو[1] [۲] مراجعه کرد.

۱،۱،۲ انتخاب پارامتر انقباضی γ و پارامتر کنترل k

برای انتخاب بهینه پارامتر انقباضی γ و پارامتر کنترل k می‌توان آماره اعتبارسنجی متقابل تعمیم یافته (GCV[2]) را به کار برد که آن را کرایون و واهبا [3] [۴] ارائه کردند. برای پارامترهای $\gamma \geq 1$ و $k \geq 0$، این معیار به صورت

$$G\ CV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_{i,k}}{1-\frac{p(k)}{n}}\right)^2 \quad (۱)$$

تعریف می‌شود، که در آن $p(k) = trace(X(X^TX + k\Omega^-)^{-1}X^T) - n_0$، n_0 تعداد $\hat{\beta}_j$ های صفر، Ω^- معکوس تعمیم یافته، $\Omega = diag(\frac{\gamma}{2}|\hat{\beta}_j|^{\gamma-2})$ و $e_{i,k} = y_i - \hat{y}_i^{Bridge}$ است.

پارامترهای γ و k چنان انتخاب می‌شوند که آماره GCV مینیمم مقدار خود را داشته باشد. بدین منظور، با در نظر گرفتن مقادیری از γ و k با فاصله‌های مساوی در ناحیه $\gamma \geq 1$ و $k \geq 0$، مقدار GCV برای هر زوج (γ, k) محاسبه می‌شود و مقداری از γ و k به‌عنوان برآورد این پارامترها انتخاب می‌شود که به‌ازای آن GCV مینیمم شده باشد [۲].

۲،۲ برآوردگر لارس

رگرسیون لارس روش خطی است که براساس الگو گرفتن از روش لاسو و روش پیشرو[4] ابداع شده است. در این روش، متغیرهای تبیینی در صورتی که معیار لازم برای ورود به مدل را داشته باشند تک به تک وارد مدل می‌شوند و بعد از ورود حذف نمی‌شوند. این روش نقص‌های روش‌های کلاسیک و پیچیدگی محاسبات روش لاسو را برطرف می‌کند.

۱،۲،۲ الگوریتم لارس

ابتدا تمام ضرایب مدل برابر صفر فرض می‌شود. متغیری که بیشترین مقدار همبستگی را با متغیر پاسخ دارد انتخاب می‌شود (x_1)، سپس بزرگترین گام ممکن در راستای متغیر تبیینی (x_1) برداشته می‌شود، به گونه‌ای که میزان همبستگی متغیر تبیینی دیگر (x_2) با باقیمانده‌های حاصل از برازش مدل برای (x_1) یکسان باشد. گام بعدی در راستای نیمساز زاویه بین (x_1, x_2) برداشته می‌شود، تا اینکه متغیر تبیینی سومی وارد مجموعه‌ای شود که به همین ترتیب بیشترین همبستگی را با مانده مدل حاصل از برازش مبتنی بر دو متغیر تبیینی اول دارد و به همین ترتیب، روند ورود متغیرها به مدل ادامه می‌یابد. در این صورت، برآورد حاصل از الگوریتم لارس $(\hat{\mu})$ به صورت

$$\hat{\mu} = X\,\hat{\beta}$$

است، که در آن $\hat{\beta}$ بردار ضرایب رگرسیونی و X ماتریس مقادیر متغیرهای تبیینی است. در هر مرحله یک متغیر تبیینی به مدل مورد نظر اضافه شده و پس از m مرحله، تعداد ضریب‌های ناصفر، m تا خواهد بود([۹] و[۱۰]).

[1] Fu
[2] Generalized Cross Validation
[3] Wahba
[4] Forward Stagewise

۳. مطالعه شبیه‌سازی

در این بخش با استفاده از فرایند شبیه‌سازی مونت کارلو به ارزیابی عملکرد و مقایسه روش کمترین توان‌های دوم عادی و لارس با بریج (با تمرکز بر روی برآوردگرهای ریج و لاسو) در مدل رگرسیون چندگانه پرداخته خواهد شد. از آن‌جا که روش‌های انقباضی برای مقابله با همخطی چندگانه است، مدل‌های شبیه‌سازی شده را بیشتر در این حالت مورد مطالعه قرار می‌دهیم. مقدار همخطی چندگانه در ماتریس‌های طرح، متفاوت است. برای داشتن داده‌هایی با میزان همخطی مختلف، مک دونالد و گلارنیو [۶] پیشنهاد دادند که داده‌ها به صورت

$$x_{ij} = (1-h^2)^{1/2} z_{ij} + h z_{ip} \qquad i = 1, \dots, n, \qquad j = 1, \dots, p$$

تولید شوند، که در آن z_{ij} اعداد شبه‌تصادفی مستقل از توزیع نرمال استاندارد و h همبستگی بین هر دو متغیر تبیینی است.

در ادامه به منظور مقایسه روش‌های رگرسیونی، شبیه‌سازی‌ها در چند وضعیت انجام شده است.

حالت ۱: نیوهاوس و اومن[1] [۷] بیان کردند در صورتی که میانگین توان‌های دوم خطا، تابعی از β، σ^2 و k باشد و متغیرهای تبیینی ثابت باشند، اگر β بردار ویژه نرمال شده متناظر با بزرگترین مقدار ویژه ماتریس X^TX تحت قید $\beta^T\beta = 1$ باشد، آن‌گاه میانگین کمترین توان‌های دوم خطا (MSE) مینیمم می‌شود.

لذا از بردار ویژه متناظر با بزرگترین مقدار ویژه ماتریس X^TX برای برآورد ضرایب بردار β به ازای $n = 5,20,50,100$، $h = 0.7,0.9,0.99$، $p = 4,20$ و $\sigma = 0.5$ در شبیه‌سازی مقادیر متغیر پاسخ استفاده می‌شود.

در تمام موارد فوق، مقدار k از روش GCV به دست آمده است. بر اساس این روش، مقداری از k انتخاب می شود که به ازای آن GCV مینیمم شود.

در همه موارد تعداد تکرار برابر ۱۰۰۰ در نظر گرفته شده و MSE نتایج حاصل از شبیه‌سازی‌ها در جدول ۱ ارائه شده‌اند.

جدول۱: نتایج شبیه‌سازی حالت۱ ($p = 4,20 \quad n = 5,20,50,100 \quad h = 0.7,0.9,0.99$)

P=20				*p=4*				*MSE*	*h*
100	50	20	5	100	50	20	5	n	
0.149	0.471	NA	NA	0.023	0.043	0.150	3715	OLS	*0.7*
0.072	0.104	0.284	0.421	0.027	0.040	0.080	0.473	Ridge	
0.159	0.483	1.282	3.617	0.021	0.440	0.144	0.850	Lasso	
0.190	0.430	1.022	3.100	0.021	0.40	0.134	0.800	Lars	
0.465	1.291	NA	NA	0.062	0.126	0.407	5135	OLS	*0.9*
0.063	0.092	0.282	0.254	0.045	0.061	0.109	0.371	Ridge	
0.470	1.025	1.942	5.322	0.062	0.125	0.342	1.363	Lasso	
0.247	0.803	0.952	4.220	0.041	0.110	0.320	1.330	Lars	
5.691	15.31	NA	NA	0.133	0.274	0.876	4955	OLS	*0.99*
0.045	0.072	0.133	0.149	0.044	0.060	0.109	0.351	Ridge	
1.612	2.129	3.451	7.474	0.133	0.255	0.566	1.751	Lasso	
1.352	1.712	2.121	5.124	0.100	0.215	0.327	1.631	Lars	

[1] Newhouse and Oman

همان‌طور که ملاحظه می‌شود با افزایش h، روش کمترین توان‌های دوم عادی نمی‌تواند ضرایب را به درستی برآورد کند اما دو برآوردگر لاسو و لارس بهتر عمل می‌کنند. همچنین برآوردگر ریج دارای بهترین عملکرد نسبت به سایر برآوردگرها است. این نتیجه زمانی که تعداد متغیرها خیلی زیاد می‌شود نیز صادق است. همچنین مشاهده می‌شود با افزایش n، همه برآوردگرها بهبود می‌یابند. بنابراین درحالت کلی وقتی تعداد متغیرها بیش از مشاهدات است یا داده‌ها دارای همخطی چندگانه باشند، برآوردگر کمترین توآنهای دوم نمی‌تواند به خوبی ضرایب را به دست آورد و مقدار میانگین توان دوم خطای آن بسیار زیاد بوده اما برآوردگر لاسو عملکرد بهتری از برآوردگر کمترین توآنهای دوم عادی دارد و دارای خطای کمتری است. همچنین برآوردگر لارس بهتر از دو برآوردگر قبلی است. در نهایت، رگرسیون ریج دارای بهترین دقت پیش‌بینی در بین سایر روش‌هاست.

حالت ۲: در این حالت بردار β به‌صورت $\beta = (3,1.5,0,1,0,0,0,0)^T$ در نظر می‌شود. در جداول ۲ و ۴ به ترتیب نتایج شبیه‌سازی برای برآورد ضرایب رگرسیونی برای n=20 و n=50 به روش‌های رگرسیون کمترین توآنهای دوم عادی، لاسو، ریج و لارس ارائه شده است.

جدول۲: نتایج شبیه‌سازی حالت ۲ $(p = 8, n = 20, h = 0.7)$

β	$\hat{\beta}_{OLS}$	$\hat{\beta}_{lasso}$	$\hat{\beta}_{Ridge}$	$\hat{\beta}_{Lars}$
3.00	0.72	1.00	1.70	1.10
1.50	-8.66	1.10	2.08	1.30
0.00	-12.40	0.90	2.16	2.61
1.00	0.90	0.00	0.33	0.15
0.00	-4.13	0.00	0.40	0.00
0.00	-7.30	0.00	0.25	-0.67
0.00	11.33	4.17	0.88	7.45
0.00	10.96	0.00	0.35	0.00

در جدول ۳، به مقایسه خطای برآوردگرهای حاصل از شبیه‌سازی روش‌های رگرسیونی حالت ۲ (با معیار میانگین توآنهای دوم خطا) پرداخته شده است.

جدول۳: میانگین توآنهای دوم خطا شبیه‌سازی حالت۲ $(p = 8, n = 20, h = 0.7)$

$\hat{\beta}_{OLS}$	$\hat{\beta}_{Ridge}$	$\hat{\beta}_{lasso}$	$\hat{\beta}_{Lars}$
355.71	40.84	62.14	58.98

در جدول ۴، نتایج شبیه‌سازی برای حالت ۲ با اندازه نمونه $n = 50$ ، ارائه شده است:

جدول۴: نتایج شبیه‌سازی حالت ۲ $(p = 8, n = 50, h = 0.7)$

β	$\hat{\beta}_{OLS}$	$\hat{\beta}_{lasso}$	$\hat{\beta}_{Ridge}$	$\hat{\beta}_{Lars}$
3.00	-1.82	2.10	2.81	2.50
1.50	3.55	1.01	1.20	1.85
0.00	4.26	0.03	0.02	0.01
1.00	7.25	2.02	1.03	0.92
0.00	1.19	0.00	0.02	0.00
0.00	-0.19	0.00	0.01	0.00
0.00	1.34	0.00	0.02	0.00
0.00	-5.95	-5.95	0.00	-0.10

برای مقایسه خطای برآوردگرهای رگرسیونی در حالت ۲ برای $n = 50$، میانگین توان‌های دوم خطای نتایج شبیه‌سازی، به صورت جدول ۵ حاصل شده است.

جدول ۵: میانگین توآنهای دوم خطا برای شبیه‌سازی حالت ۲

$\hat{\beta}_{OLS}$	$\hat{\beta}_{Ridge}$	$\hat{\beta}_{lasso}$	$\hat{\beta}_{Lars}$
93.58	17.21	23.93	22.11

جداول ۲ و ۴ نشان می‌دهند وقتی مقدار واقعی پارامتر β صفر و میزان همخطی چندگانه قابل توجه است، برآوردگر کمترین توان‌های دوم نمی‌تواند ضرایب را به درستی پیدا کرده و مقدار MSE آن بسیار بالا است. اگرچه برآوردگر ریج دارای MSE کمتری است. با توجه به جدول ۲ و۴ برآوردگرهای لاسو و لارس در برآورد پارامترهای صفر عملکرد خوبی دارند و برخی از ضرایب متغیرهای بی اثر دقیقاً صفر برآورد می شوند و به این ترتیب متغیرهای مربوط به آن ضرایب از مدل خارج می‌شوند. با توجه به مقدار کمتر MSE در روش لارس نسبت به لاسو و امکان برآوردیابی و انتخاب متغیرها به طور همزمان این روش از مزایای به کارگیری آن است.

به طور کلی نتایج شبیه‌سازی نشان می‌دهد که روش‌های انقباضی ریج، لارس و لاسو عملکرد بهتری نسبت به رگرسیون کمترین توان‌های دوم معمولی دارند و در صورت بروز همخطی و در مسائل بعد بالا که تعداد مشاهدات نسبت به تعداد متغیرها کم است، به کارگیری روش‌های لارس و ریج توصیه می‌شود.

۱,۳ کاربرد روشهای رگرسیونی در دادههای رشد گندم

داده‌های آیووا (IOWA) میزان رشد گندم برای ایالت آیووا در سال‌های ۱۹۳۰-۱۹۶۲ را با توجه به میزان بارندگی فصل قبل، سه ماه فصل رشد، فصل برداشت و همچنین متوسط درجه حرارت اندازه‌گیری شده برای این پنج ماه را نشان می‌دهد. در جدول۶ توصیفی از این متغیرها و در جدول ۷ آماره‌های توصیفی متغیر پاسخ (Yield) آمده است.

جدول۶: توصیف متغیرهای دادههای آیووا

نوع متغیرها	توصیف	متغیرها
متغیر پاسخ	میزان گندم داده شده سال	Yield
متغیر تبیینی	سال اندازه‌گیری	Year
" "	بارش فصل قبل	Rain0
" "	متوسط دمای اولین ماه رشد	Temp1
" "	متوسط بارش اولین ماه رشد	Rain1
" "	متوسط دمای دومین ماه رشد	Temp2
" "	متوسط بارش دومین ماه رشد	Rain2
" "	متوسط دمای سومین ماه رشد	Temp3
	متوسط بارش سومین ماه رشد	Rain3
	متوسط دمای ماه برداشت	Temp4

آماره‌های توصیفی متغیر پاسخ در داده‌های آیووا در جدول ۷ آمده است:

جدول۷: آماره‌های توصیفی متغیر پاسخ در دادههای آیووا

مینیمم	چارک اول	میانه	چارک سوم	ماکسیمم	میانگین	انحراف معیار
20.0	43.1	52.0	59.0	76.0	50.0	13.18

برآورد ضرایب برآوردگرهای پارامترهای مدل رگرسیونی برای این داده‌ها به چهار روش رگرسیون OLS، لاسو، ریج و لارس در جدول ۸ ارائه شده است:

جدول۸: برآورد پارامترها برای داده‌های آیووا همراه با خطای آنها

	$\hat{\beta}_{OLS}$	$\hat{\beta}_{lasso}$	$\hat{\beta}_{Ridge}$	$\hat{\beta}_{Lars}$
	0.00	0.00	0.00	0.66
	0.88	0.76	0.62	0.00
	0.78	0.28	0.43	0.00
	-0.46	0.00	-0.29	0.00
	-0.78	0.00	-0.77	0.00
	0.48	0.00	0.12	1.69
	2.56	2.10	2.08	-0.17
	0.05	-0.01	-0.05	0.00
	0.41	0.24	0.64	-0.23
	-0.66	-0.55	-0.51	0.66
MSE	61.83	50.62	47.49	42.57

جدول ۸ نشان میدهد که برای برآورد میزان گندم بر اساس ۹ متغیر تبیینی، روش لارس دارای بهترین عملکرد است و همچنین تعداد ضرایب صفر برآورد شده آن بیش از سایر روش هاست که موجب سادگی بیشتر مدل می شود. این توانایی روش لارس در نتایج شبیه‌سازی نیز ملاحظه شد. روش ریج نیز پس از لارس دارای عملکرد بهتری نسبت به سایر روش هاست و رگرسیون لاسو در مرتبهٔ بعد از آن قرار دارد. روش کمترین توان‌های دوم عادی دارای بیشترین خطا بوده است.

نتیجه‌گیری

بهبود دادن برآوردگرهای معادلات رگرسیون زمینه کاری بسیاری از آماردانان دهه اخیر است. یکی از عوامل که منجر به تلاش برای یافتن برآوردگرهای بهبودیافته می‌شود، وجود همخطی در ماتریس طرح است. همان‌طور که در این مقاله شرح آن گذشت، یکی از این برآوردگرهای بهبودیافته که در سال های اخیر از سوی آماردانان مورد توجه زیادی قرار گرفته است، برآوردگر بریج و حالات خاص آن یعنی ریج و لاسو است. وقتی با مسأله همخطی مواجه هستیم، برآوردگر بریج در رده برآوردگرهای اریب، بهترین برآوردگر است و خطای آن از خطای برآوردگر کمترین توان‌های دوم عادی کمتر است و همچنین دارای ویژگی‌های مفید و خوبی است که استفاده از این برآوردگر را توجیه می‌کنند.

در این مقاله، به برآورد ضرایب معادلات رگرسیونی چندگانه با در نظر گرفتن مساله همخطی چندگانه و استفاده از داده‌هایی که در آنها این مشکل به چشم می‌خورد، پرداخته شده است. همچنین برآوردگر بریج مانند ریج و لاسو را با برآوردگر لارس مورد بررسی قرار داده و دریافتیم که برآوردگر لارس نیز در مقابله با پدیدهٔ همخطی چندگانه بسیار خوب عمل کرده است و همچنین در این روش، برآوردیابی و انتخاب متغیر تواماً صورت می‌پذیرد که این ویژگی این برآوردگر سبب محبوبیت آن در مسائل با بعد بالا شده است.

به عنوان ارائه کاربردی از نتایج حاصل، داده‌های برداشت گندم به کار گرفته شد. برازش روش های رگرسیونی مورد نظر به این داده‌ها نیز نتایج حاصل از شبیه‌سازی ها را تایید می کند.

References

1. Tibshirani R., “Regression Shrinkage and Selection via the lasso”, Journal of Royal Statistical Society (1966), 58,1,267-88.
2. FU, J., “Penalized Regressions: The Bridge Versus the Lasso”, Journal of Computational and Graphical Statistics, Vol. 7, No. 3, (1998), 397-416.
3. Frank I. E. and Friedman J. H., “ A statistical view of some chemometrics regression tools (with discussion)”, Technom., 35, (1993), 109–148.
4. Craven, P. and Wahba, G., "Smoothing Noisy Data With Spline Functions", Numerische Mathematik,31, (1979), 377–403.
5. James W. and Stein C., “Estimation with quadratic loss”, Proceeding of the forth berkely symposium, 1, (1961), 361-379.
6. Mc. Donald. G. C and Galarneau. D. I. ,“A monte carlo evaluation of some ridge type estimators", Amer. Statist. Assoc. (1975), 70 , 407-416.

7. Newhouse. J. P. and Oman. S. D., "An evaluation of ridge estimators", Rand Corporation, P-716-PR
8. Mallick, H. and Yi, N., "Bayesian Bridge Regression", Journal of Applied Statistics, Vol. 45, No. 6, (2018), 988-1008.
9. Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R., "Least Angle Regression", The Annals of Statistics, Vol. 32, No. 2, (2004) 407-499.
10. Hesterberg, T., Choi, N. H., Meier, L. and Fraley C., "Least Angle and L_1 Penalized Regression: A review", Statistical Surveys, (2008), Vol. 2, 61-93.