فصلنامه پژوهش‌های کاربردی روان‌شناختی
۱۳۹۹, ۱۱(۳), ۱۹۳-۲۰۶
مؤسسه روان‌شناسی و علوم تربیتی
دانشکده روان‌شناسی و علوم تربیتی
دانشگاه تهران

Journal of Applied Psychological Research
2020, 11(3), 193-206
Institute of Psychology and Educational Sciences
Faculty of Psychology and Educational Sciences
University of Tehran

(DOI): 10.22059/japr.2020.305766.643568

مقایسه رگرسیون خطی و الگوریتم‌های رگرسیون انقباضی (ستیغی، لسو و الستیک شبکه‌ای) با استفاده از داده‌های بیماران استرس پس از سانحه

Comparing Linear Regression to Shrinkage Regression Algorithms (RR, Lasso, El Net) Using PTSD Patients' Data

حجت‌اله فراهانی*

Hojjatollah Farahani

چکیده

هدف از این پژوهش، معرفی الگوریتم‌های رگرسیونی جایگزین، برای رگرسیون خطی بود. به‌این‌منظور، از الگوریتم های رگرسیونی نوین چونان ستیغی، لسو والستیک استفاده شد که در آنان دقت پیش‌بینی از رهگذر میزان‌سازی تابع هزینه بیشینه می‌شود. در این پژوهش، نخست به توضیح مبانی نظری این الگوریتم‌های نوین پرداخته شد و سپس در قالب یک مثال عددی با استفاده از داده‌های بیماران استرس پس از سانحه به تفسیر خروجی نرم‌افزار و مقایسه آن‌ها مبادرت شد. جامعه پژوهش افراد مبتلا استرس پس از سانحه شهر تهران در سال ۱۳۹۹ بود که داده‌های ۹۷ بیمار (۷۳ زن و ۲۴ مرد) که در آن‌ها هشت متغیر مرتبط با شدت بازتجربه ترما اندازه‌گیری شده بود به‌عنوان نمونه پژوهش بررسی شدند. داده‌ها با رگرسیون چندگانه خطی کلاسیک، رگرسیون ستیغی (RR)، لسو (Lasso) و رگرسیون الستیک شبکه‌ای (Elastic) با استفاده از نرم‌افزار R تحلیل شدند. یافته‌ها نشان داد که رگرسیون الستیک، لسو و ستیغی، به‌ترتیب بیشترین درصد واریانس تبیینی و کمترین میانگین خطا را در مقایسه با رگرسیون خطی نشان دادند. در شرایطی که مفروضه‌های نبود هم‌خطی و ثبات واریانس باقیمانده‌ها احراز نشود، کاربست رگرسیون خطی مشکل‌ساز و کاربرد این روش‌های جایگزین پیشنهاد می‌شود.

واژه‌های کلیدی: عدم و ثبات واریانس، هم‌خطی چندگانه، رگرسیون ستیغی، رگرسیون لسو، رگرسیون شبکه‌ای الستیک

Abstract

The purpose of this research was to introduce the alternative model of regression algorithms and having it compared to linear regression. To do this, we need to use modern algorithms such as Ridge, Lasso, and Elastic net regression in which precision is maximized by regularizing the cost function. In this paper theoretical basis and practical implications have been explained.. The target population was patients diagnosed with Post Traumatic Stress Disorder (PTSD) in 2020 for the comparison. 97 PTSD patients (73 females and 24 males) in Tehran were measured in 8 variables related with the intensity of the trauma re-experience. The linear regression, Ridge, Lasso, and Elastic Regressions were used with R software. The results indicated that compared to linear regression, Elastic. LASSO and Ridge explained more variances and had more R square and less MSE respectively. When the main assumptions of Linear regression are not met, using shrinkage regressions seems to be reasonable and accurate.

Keywords: Penalized Algorithms, Ridge Regression, Lasso Regression, Elastic Net Regression, Multi-Collinearity, Heteroscedasticity

***نویسنده مسئول:** استادیار گروه روان‌شناسی، دانشکده علوم انسانی، دانشگاه تربیت مدرس، تهران، ایران

Email: h.farahani@modares.ac.ir

دریافت: ۹۹/۰۳/۱۶ **پذیرش:** ۹۹/۰۶/۰۸

Received: 6 Jun 2020 **Accepted:** 29 Aug 2020

مقدمه

مدل‌سازی‌های مبتنی بر رگرسیون کمترین مربعات متداول[1] با هدف تبیین (پیش‌بینی) متغیر وابسته در پژوهش‌های علوم روان‌شناختی گستردگی و ژرفای چشمگیری یافته است (فکس، ۲۰۱۶). رگرسیون خطی، به‌عنوان مجموعه‌ای از روش‌های آماری نیز مانند دیگر روش‌ها، مستلزم در نظرگیری مفروضه‌هایی است که درک نادرست از آن مفروضه‌ها و شرایط کاربست این روش گسترده و گمراه‌کننده است. ردپای این کاربست نادرست در بسیاری از پژوهش‌های داخلی نیز چشم‌آزار و ذهن‌کاه است. بنابراین، به‌نظر می‌رسد که باید در مورد کاربست این روش و نیز روش‌های نوین و جایگزین به‌گونه‌ای جدی چاره اندیشیده شود. مشکل اصلی در پژوهش‌هایی که از روش رگرسیون خطی استفاده کرده‌اند، این است که در بسیاری از موارد مفروضه‌های مهم خطی‌بودن رابطه، عدم وجود هم‌خطی بین پیش‌بین‌ها، کفایت حجم نمونه، توازن نسبت تعداد پیش‌بین‌ها به تعداد افراد نادیده گرفته شده است. در پژوهش ارنس و البرس (۲۰۱۷) در ۲۹ درصد از موارد کاربرد رگرسیون در مقاله‌های با کاربرد رگرسیون در پژوهش‌های روان‌شناسی بالینی، مفروضه‌ها پیش از کاربرد وارسی نشده‌اند. هدف از پژوهش حاضر، این است که جایگزین‌های مناسب رگرسیون خطی و شرایط و چگونگی اجرا و تفسیر نتایج آن‌ها توضیح داده شود که به‌گاه ضرورت، روان پژوهشگران از آن‌ها در مدل‌سازی پژوهش‌های خود بهره‌ای گیرند. نخست به بررسی این روش‌ها از نظرگاه نظری پرداخته می‌شود و آن هنگام که سبد ذهن پژوهشگر از این مبانی آکنده گشت به کاربست و مقایسه آن‌ها پرداخته می‌شود.

رگرسیون خطی چندگانه

رگرسیون خطی چندگانه[2]، یکی از پرکاربست‌ترین روش‌ها در مدل‌سازی‌های علوم روان‌شناختی است که برای نشان دادن تأثیر چندین متغیر پیش‌بین بر یک نتیجه کمّی پیوسته (متغیر ملاک) کاربرد دارد.
مدل آماری این روش به‌گونه زیر است:

$$Y_i=\beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots+\beta_kX_k+\alpha$$

در این معادله می‌توان مقدار موردی از y_i را با مقادیر X_{ij} در متغیرهای مستقل $(j=1\cdots P)$ و X_i پیش‌بینی کرد.

شیب‌های (β_j) متفاوت هرکدام نشان‌دهنده میزان رابطه بین متغیر مستقل X_j و وابسته Y است.

در پژوهش‌های همبستگی گاهی به متغیر مستقل، پیش‌بین و به متغیر وابسته، ملاک گفته می‌شود.

خطای یک y_i معین، برابر با تفاوت بین مقدار مشاهده شده و مقدار پیش‌بینی شده‌ای حاصل می‌شود که مدل رگرسیونی آن را به‌دست می‌دهد، مقدار خطا را با ε_i نشان می‌دهند که فرض بر آن است که با مقادیر X_P نامرتبط است. $\beta_0(\alpha)$ که عرض از مبدأ یا مقدار ثابت نامیده می‌شود؛ مقدار y، به ازای همه مقادیر پیش‌بین صفر در مدل است. در مدل رگرسیون چندگانه P، متغیر پیش‌بین وجود دارد و آن‌گاه که P=۱ شود

1. Ordinary Least Squares (OLS)
2. linear multiple regression

آن را رگرسیون خطی ساده می‌نامند (مونتمگری و پک‌وینینگ، ۲۰۱۲).

چهار مفروضه اصلی رگرسیون خطی چندگانه

باید به‌خاطر داشت، مدل رگرسیون، فقط آن هنگام استنباط مناسبی را به‌دست می‌دهد که مفروضه‌های آن درست باشند؛ هرچند این مدل نسبت به تخطی ملایمی از این مفروضه‌ها تا حدی مقاوم[1] است. منابع آماری متعددی از جمله کوهن، کوهن، وست و ایکن (۲۰۰۳)؛ هیر، ببین، اندرسن و بلک (۲۰۱۸)؛ مونتمگری و پک وینینگ (۲۰۱۲)؛ فراهانی و عریضی (۱۳۸۷) پیشینه گسترده و بیشتری از چنین مفروضه‌هایی همراه با چگونگی مقابله با تخطی از آن‌ها فراهم آورده‌اند.

تخطی از این مفروضه‌ها به انواع متنوعی از مشکل می‌انجامد. نخست، برآورد سوگیرانه[2] و بی‌ثبات[3] (نایکنواخت) می‌شود و دیگر این‌که اگر برآورد مبتنی بر کمترین مربعات متداول[4] (OLS) باشد که بیشتر مواقع چنین است، ممکن است در این صورت دیگر این روش کارساز نباشد و آخر آن‌که آزمون معناداری فرضیه صفر و فاصله‌های اطمینان ممکن است نادرست باشند، بدین‌معنا که مقادیر P به‌طور منظم بسیار کوچک یا بسیار بزرگ و فاصله‌های اطمینان بسیار باریک یا گسترده می‌شوند (ویلیامز، گراجلز و کرکویکز (۲۰۱۳)؛ کوهن و همکاران (۲۰۰۳).

به‌خاطر داشته باشید این مفروضه‌ها، تنها در مورد رگرسیون با روش OLS درست است که به‌طور متداول و پیش‌انتخاب در نرم‌افزارهای آماری موجود است و در مقالات روان‌شناسی به غایت کاربست یافته است و در مورد دیگر روش‌ها درست نیست، مفروضه‌های چهارگانه خطی‌بودن، نرمالیتی، ثبات واریانس و استقلال حائز اهمیت‌اند که جزئیات آن در منابع مختلف آماری و از جمله هاول (۲۰۱۳)؛ هیر و همکاران (۲۰۱۵)؛ مونتمگری و پک‌وینینگ (۲۰۱۲)؛ فراهانی و عریضی (۱۳۸۷) قابل‌مطالعه است. دو شرط وجود هم‌خطی چندگانه و نرمال بودن نمرات باقیمانده بسیار مشکل‌ساز است.

مدل‌های رگرسیونی به شیوه‌های مختلفی انجام می‌شوند. دست‌کم ۱۶ روش برای انجام وجود دارد.

افزون بر مشکلات گفته شده در کاربست رگرسیون چندگانه خطی در پژوهش‌های علوم روان‌شناختی، پژوهشگران با مدل‌سازی‌هایی روبه‌رو می‌شوند که در آن‌ها میزان‌سازی[5] صورت می‌گیرد که در قالب این میزان‌سازی برآورد پارامترها دقیق‌تر و باثبات‌تر می‌شود (سمکر و الپو، ۲۰۱۰).

هدف از این میزان‌سازی، نوعی مقیاس‌سازی بهینه است. میزان‌سازی برای اصلاح مدل‌های رگرسیونی به‌کار می‌رود. در میزان‌سازی تلاش می‌شود تا با استفاده از الگوریتم‌های انقباضی مقادیر b را مناسب سازند. انواع رایج از الگوریتم‌های انقباضی رگرسیون وجود دارد که عبارتند از:

---

1. robust
2. biased
3. inconsistent
4. ordinary least squares
5. regularization

۱- رگرسیون ستیغی[1] (RR)؛
۲- رگرسیون لسو[2] (Lasso) و
۳- رگرسیون شبکه الاستیک[3] (El net).

در ادامه به بررسی مبانی و نکات این روش‌ها به‌عنوان روش‌هایی جایگزین رگرسیون چندگانه خطی پرداخته می‌شود.

۱- رگرسیون ستیغی

رگرسیون ستیغی، زمانی بسیار مناسب است که تعداد زیادی متغیر پیش‌بین وجود دارد که ضرایب آن‌ها غیرصفر است و از توزیع نرمال استخراج شده باشند. به‌ویژه، این روش زمانی بسیار مناسب است که تعداد متغیرهای موجود در مدل زیاد و یا هم‌خطی چندگانه شدید وجود داشته باشد. در این هنگام واریانس برآوردگرها متورم و به شکل قله (ستیغ) خود را نشان می‌دهد. یکی از مشکلات اصلی در رگرسیون چندگانه خطی، این است که تعداد زیاد متغیرهای پیش‌بین سبب بیش‌برازندگی[4] و تعداد کم متغیرهای پیش‌بین سبب کم‌برازندگی[5] می‌شود؛ بنابراین برای چیرگی بر این مشکل باید راهی برای تعیین تعداد مناسب برآوردگرها یافت که برآوردگرها نااریب بوده و واریانس کوچک‌تری نسبت به OLS داشته باشند، در این حالت رگرسیون ستیغی مناسب است (ویلکاکس، ۲۰۱۹).

در روش رگرسیون خطی هدف کمینه‌سازی مجموع مربعات خطا است

$$\arg\min \| y - \hat{y} \|^2 = \arg\min \sum [y_i - (B_0 + B_1 X_1) + \beta_2 X_2 + \cdots + \beta X_P)]^2$$

این هدف در شرایطی که تعداد متغیرهای پیش‌بین افزایش باید یا هم‌خطی چندگانه آن‌ها وجود داشته باشد مخدوش می‌شود. محصول چنین خدشه‌ای بیش برازندگی و افزونگی هم‌خطی است.

برای حل این مشکل از رگرسیون ستیغی استفاده می‌شود. مجموع مربعات خطای رگرسیون ستیغی عبارت است از:

$$\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{i=1}^{n} x_{ij}\beta_j)^2 + \lambda \sum_i \beta_i^2$$

در این معادله λ میزان جریمه[6] است که به تعداد پارامترها برمی‌گردد. در رگرسیون خطی چندگانه λ صفر است. هر چقدر مقدار λ بزرگ‌تر باشد، مقدار انقباض بیشتری صورت می‌گیرد. منظور از argmin مقادیری از

1. Ridge Regression (RR)
2. Leasd Absolute Shrinkage and Selection Operator (LASSO)
3. Elastic Net (EL Net)
4. over fitting
5. under fitting
6. penalty

β هستند که تابع موردنظر را کمینه می‌کنند (مرونا، ۲۰۱۱). یکی از ویژگی‌های رگرسیون ستیغی این است که جریمه، ضریب را به‌سمت صفر می‌کشاند؛ اما هیچ‌کدام را دقیقاً صفر نمی‌کند، مگر آن‌که λ بسیار بزرگ باشد (لاکمن، ارولو و ای‌ایند، ۲۰۱۴).

یکی از مشکلات اصلی در رگرسیون ستیغی، تعیین λ است؛ البته می‌توان با ترسیم نمودار MSE برحسب λ نیز پاسخ درستی را یافت یا از طریق اعتباریابی متقابل[1] آن را به‌دست آورد. در این مقاله روش دوم به‌کار گرفته شده است.

با توجه به جمیع جهات، خلاصه مفاهیم و کاربردها در رگرسیون ستیغی به‌صورت زیر است:

* رگرسیون چندگانه (OLS) بهترین پیش‌بینی را در شرایطی فراهم می‌سازد که مفروضه‌های ضروری در تدوین مدل احراز شده باشد.
* در رگرسیون ستیغی به ویژگی‌های مهم‌تر وزن بیشتری داده می‌شود.
* از مقدار R^2 یا MSE می‌توان در تعیین مقدار مناسب λ استفاده کرد.
* هنگام استفاده از RR باید داده‌ها را ابتدا استاندارد کرد (سله، ارشی و کبریا، ۲۰۱۹).

۲- رگرسیون Lasso (عملگر گزینش و انقباض کمترین قدر مطلق)

هدف از این رگرسیون به‌عنوان جایگزین خطی (OLS)، این است که دقت پیش‌بینی را بهبود داده و امکان تفسیر مدل را با ایجاد زیرگرد آیه‌های کوچک‌تری از متغیرهای کمکی با بیشترین اثر را فراهم آورد. در این الگوریتم، رگرسیونی مجموع مربعات خطای رگرسیون به‌صورت زیر است:

$$\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{n}x_{ij}\beta_j)^2+\lambda\sum_{j}|\beta_i^2|$$

در این معادله λ پارامتر میزان‌سازی مدل است. اگر مقدارش برابر با صفر باشد، مدل به‌OLS تبدیل می‌شود. با افزایش آن تعداد متغیرهای مستقل در مدل کاهش می‌یابند.

در مقایسه با روش RR، هر دو روش مشکل هم‌خطی چندگانه را حل می‌کنند و مدلی بدون بیش‌برازندگی یا کم‌برازندگی ارائه می‌کنند، روش RR نسبت به Lasso سریع‌تر است (لیو و ژنگ، ۲۰۰۹).

برای تعیین λ در هر دو روش، از راه‌های مختلف از جمله اعتباریابی متقابل استفاده می‌شود و متغیرها هم باید مانند RR استاندارد شوند (حستی، تبشیرانی و وینرایت، ۲۰۱۵).

۳- رگرسیون شبکه الستیک (El net)

در این الگوریتم رگرسیونی لسو و ستیغی، تلفیق می‌شود و جایگزین آن‌ها است. در نتیجه میزان‌سازی مرتبه ۱ و ۲ روی مدل هم‌زمان انجام می‌شود:

1. k-fold cross validation

$$\min(\sum y_i - (\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)^2+\lambda_1\sum\beta_i^2+\lambda_2\sum|B_i|)$$

رگرسیون شبکه الستیک نیز، زمانی کاربرد و مناسبت دارد که هم‌خطی چندگانه وجود داشته باشد و این روش نیز، مانند دو روش پیشین فرض نرمال‌بودن نمرات باقیمانده (خطا) ضرورت ندارد (پارک و کانیشی، ۲۰۱۵).

می‌توان گفت، هدف از رگرسیون‌های انقباضی ایجاد تعادل بین سادگی[1] و درستی[2] مدل است. متغیریابی و پیش‌بینی متغیرها از داده‌ها با ابعاد بسیار بالا مانند امواج مغزی ممکن است چالش‌برانگیز باشد. به‌ویژه در مواقعی که حجم نمونه بسیار کوچک است و تعداد متغیرهای پیش‌بین زیاد ($n<<p$). در چنین مواقعی می‌توان از روش‌های رگرسیونی انقباضی استفاده کرد. از جمله این روش‌ها رگرسیون ستیغی، Lasso و رگرسیون شبکه الاستیک است (ژو و حستی، ۲۰۰۵).

در RR و Lasso، عبارت میزان‌سازی به تابع هزینه در رگرسیون اضافه می‌شود. این عبارت تابع هزینه رگرسیونی را جریمه می‌کند. به‌گونه‌ای که مجموع قدرمطلق مقادیر ضرایب افزایش یابد.

در Lasso این عبارت نرم L1 نامیده می‌شود و در RR این عبارت نرم L2 یا میزان‌سازی تیکانف[3] نامیده می‌شود. نرم مبتنی از مجموع ضرایب استاندارد و فرم نسبتی از مجموع مجذور ضرایب است.

$$\lambda(\frac{1}{2}(1-\alpha)\beta^2+\alpha|\beta|)$$

در این معادله، پارامتر α نوع انقباض و پارامتر جریمه λ مقدار انقباض را تعیین می‌کند. در Lasso ($\alpha=1$) و در رگرسیون ستیغی ($\alpha=0$) است و در شبکه الاستیک پارامتر α بین مقادیر ۱ و ۰ ($0<\alpha<1$) انتخاب شود که با افزایش مقدار α عبارت جریمه افزایش می‌یابد (پارک و کانیشی، ۲۰۱۵؛ ژو و حستی، ۲۰۰۵).

در RR تنها ضرایب به‌سوی صفر شدن بدون صفرسازی آن‌ها انجام می‌شود در حالی که در Lasso برخی ضرایب خنثی (صفر) می‌شوند. روش El net بین RR و Lasso قرارداد و محدودیت‌های آن دو را ندارد. میزان‌سازی با استفاده از پارامتر λ انجام می‌شود. شدت میزان‌سازی با افزایش مقدار λ شدت می‌گیرد. روش‌های مبتنی بر انقباض ضرایب، ریشه در نظریه استین (۱۹۸۱) دارد. در این روش‌ها کوشش می‌شود تا صحت پیش‌بینی بهبود یابد. این کار با انقباض پارامترهای برآورد شده یا صفرسازی آن‌ها که واریانس را کاهش می‌دهد صورت می‌گیرد (ژو و حستی، ۲۰۰۵؛ پارک و کانیشی، ۲۰۱۵).

روش Lasso، بهترین روش در شرایطی است که موارد اندکی انتظار است که انتخاب شوند و در شرایطی که بخش بزرگی از متغیرها انتظار انتخاب شدن دارند RR بهترین روش است و El net در شرایطی است که انتظار چیزی بین آن دو است. باید توجه داشت که RR و Lasso در چگونگی مواجهه با متغیرها همبسته

1. simplicity
2. accuracy
3. Tikhonov

(هم‌خطی) متفاوت عمل می‌کنند. در RR متغیرهای همبسته به‌سوی یکدیگر منقبض می‌شوند، اما در Lasso نوعاً یکی انتخاب می‌شود؛ بنابراین RR از Lasso در شرایطی که متغیرها همبستگی بالایی دارند عملکرد بهتری دارد و انتخاب کمی زیر ۱ به مدل امکان می‌دهد تا متغیرهای پیش‌بینی بیشتری انتخاب کرد.

انتخاب پارامتر λ را می‌توان از طریق اعتبارسازی متقابل با K بلوک به‌دست آورد. در این روش، داده‌ها نخست به‌طور تصادفی به k بلوک[1] با حجم برابر تقسیم می‌شود. سپس پارامترها محاسبه و در یک بلوک برازش می‌یابد (برای داده‌هایی که در K-1 بلوک باقی می‌مانند). سپس از مدل بر ارزش یافته برای تخمین خطای پیش‌بینی در آن بلوک که کنار گذاشته شده استفاده می‌شود. این روند برای همه k بلوک استفاده می‌شود تا خطای پیش‌بینی برای تعیین شود (پارک و کانیشی، ۲۰۱۵).

ارزیابی و مقایسه مدل‌های رگرسیونی

برای نشان دادن این‌که کارکرد کدام الگوریتم انقباضی بهتر است، از شاخص R^2، تعداد متغیرهای پیش‌بین معنادار، میانگین مربعات خطا[2] استفاده می‌شود.

برای انجام مراحل اجرای الگوریتم‌های انقباضی می‌توان از نرم‌افزارهای R و Python استفاده کرد. پژوهشگران چنان‌چه بنابر علاقه یا ضرورت در پی کاربرد این روش‌ها هستند، برای دریافت کدها و مراحل لازم می‌توانند به نویسنده این مقاله ایمیل ارسال کنند.

مثال عددی

برای مقایسه سه الگوریتم انقباض رگرسیونی با رگرسیون خطی چندگانه داده‌های ۹۷ بیمار مبتلا استرس پس از سانحه شهر تهران در سال ۱۳۹۹که در آن‌ها هشت متغیر مرتبط با شدت باز تجربه ترما اندازه‌گیری شده است موردبررسی قرار گرفت. لازم به یادآوری است که صرفاً این یک مثال است و این روش‌ها در شرایط مشابه که به آن پرداخته شد توجیه‌پذیر است. ابتدا رگرسیون چندگانه خطی هم‌زمان کلاسیک انجام می‌شود، سپس رگرسیون ستیغی و Lasso و رگرسیون Elastic اجرا شد. برای تعیین λ یک راه‌حل مناسب آزمایش و خطا است به‌گونه‌ای که MSE کمینه باقی بماند و راه‌حل رایج‌تر استفاده از اعتباریابی متقابل است. جدول ۱ خروجی نرم‌افزار SPSS برای رگرسیون خطی چندگانه هم‌زمان را نشان می‌دهد.

جدول ۱- ضرایب استاندارد، غیراستاندارد، آزمون t و سطح معناداری رگرسیون خطی هم‌زمان

مدل	B	β	t	P	Tolerance	VIF
مقدارثابت	۴۹/۵۲	----	۴/۶	۰/۰۰۱	------	------
جنس	-۲۰/۴۵	-۰/۲۵	-۱/۶۷	۰/۰۹۷	۰/۳۵	۲/۹۹
وضعیت تأهل	۰/۵۸۰	-۰/۰۰۴	-۰/۰۳۱	۰/۹۷۶	۰/۲۶	۳/۸۲

1. fold
2. MSE

تحصیلات	۰/۴۲	۰/۷۴	۰/۷	۰/۴۸۷	۰/۱۷	۵/۸۸
شغل	۰/۱۲	۰/۱۱	۰/۰۹۳	۰/۹۲۶	۰/۱۴	۷/۲۴
سن	−۰/۲	−۰/۲۱	−۲/۱۹	۰/۰۳	۰/۵۱	۱/۹۶
تعدادفرزندان	۵/۱۵	۰/۱۷	۱/۶۸	۰/۰۹۷	۰/۶۱	۱/۶۴
سازش یافتگی	−۰/۳۴	−۰/۲۶	−۳	۰/۰۰۳	۰/۱۸	۵/۵۵
اضطراب	۰/۹۳	۰/۴۳	۵	۰/۰۰۱	۱۷/۰	۶/۰۲۴
متغیر وابسته: تجربه‌تروما					**R^2=۰/۳۶۵**	

همان‌گونه که نتایج این تحلیل نشان می‌دهد اضطراب، سازش یافتگی و سن به‌طور معناداری شدت بازتجربه تروما در این بیماران به‌طور معناداری پیش‌بینی می‌کند. میزان درصد تبیین واریانس ۳۶/۵ است. خروجی نرم‌افزار R مشخصات رگرسیون ستیغی و جدول ضرایب در جدول‌های ۲ و ۳ آمده است.

جدول ۲- مشخصات اجرای رگرسیون ستیغی

گروه تحلیل	گوسین
مقدارآلفا	صفر
تعداد مقادیر لامبدا	۱۰۰
استانداردسازی	انجام شده است
آستانه همگرایی	۰/۰۰۰۰۰۰۱
بیشینه از سرگیری	۱۰۰۰۰۰
تعداد بلوک برای اعتباریابی چندگانه	۱۰
داده‌های حاصل نشده	حذف شدند
مقدار λ بهینه حاصل از اعتباریابی چندگانه	۳/۴۴

در این جدول مقدار آلفا همان‌گونه که گفته شد برای رگرسیون ستیغی صفر است و مقدار λ بهینه از طریق ۱۰ بلوک اعتباریابی برابر با ۳/۴۴ به‌دست آمده است.

جدول ۳- ضرایب متغیرها در رگرسیون ستیغی

مقدارثابت	۴۶/۲۲
جنس	-۱۰/۴۲
وضعیت تأهل	-۰/۸۹
تحصیلات	۰/۱۸
شغل	۰/۰۹۰
سن	-۰/۱۵
تعداد فرزندان	۳/۲۲
سازش‌یافتگی	-۰/۲۹
اضطراب	۰/۷۲
متغیر وابسته: تجربه‌تروما	**R^2=۰/۴۲۳**

همان‌گونه که نتایج در جدول ۳ نشان می‌دهد ضرایب تغییرکرده‌اند و براساس میزان جریمه، میزان‌سازی صورت گرفته است و درصد واریانس تبیین شده افزایش یافته است. خروجی نرم‌افزار R مشخصات رگرسیون در جدول ضرایب در جدول‌های ۴ و ۵ آمده است.

جدول ۴- مشخصات اجرای رگرسیون لسو

گروه تحلیل	گوسین
مقدار آلفا	۱
تعداد مقادیر لامبدا	۱۰۰
استانداردسازی	انجام شده است
آستانه همگرایی	۰/۰۰۰۰۰۰۱
بیشینه از سرگیری	۱۰۰۰۰۰
تعداد بلوک برای اعتباریابی چندگانه	۱۰
داده‌های حاصل نشده	حذف شدند
مقدار لامبدای بهینه λ حاصل از اعتباریابی چندگانه	۰/۱۸۸

در جدول ۴، مقدار آلفا همان‌گونه که گفته شد برای رگرسیون لسو ۱ است و مقدار λ بهینه از طریق ۱۰ بلوک اعتباریابی برابر با ۰/۱۸۸ به‌دست آمده است.

جدول ۵- ضرایب متغیرها در رگرسیون لسو

مقدارثابت	۴۷/۲
جنس	-۱۶/۱۴
وضعیت تأهل	-۰/۲۳
تحصیلات	۰/۲۶
شغل	صفر
سن	-۰/۱۹
تعدادفرزندان	۴/۲۲
سازش یافتگی	-۰/۳۵
اضطراب	۰/۸۹
متغیر وابسته: تجربه تروما	**R^2=۰/۴۵۶**

همان‌گونه که نتایج در جدول ۵ نشان می‌دهد، یکی از ضرایب کوچک در جدول ۱ صفر شده است و میزان درصد واریانس تبیین شده افزایش یافته و به ۰/۴۵۶ است. براساس نتایج رگرسیون ستیغی و لسو، تابع هزینه (جریمه) رگرسیون را جریمه می‌کند به‌گونه‌ای که مجموع قدر مطلق مقادیر ضرایب افزایش یابد. خروجی نرم‌افزار R مشخصات رگرسیون الستیک در جدول ضرایب در جدول ۶ و ۷ آمده است.

جدول ۶- مشخصات اجرای رگرسیون شبکه الستیک

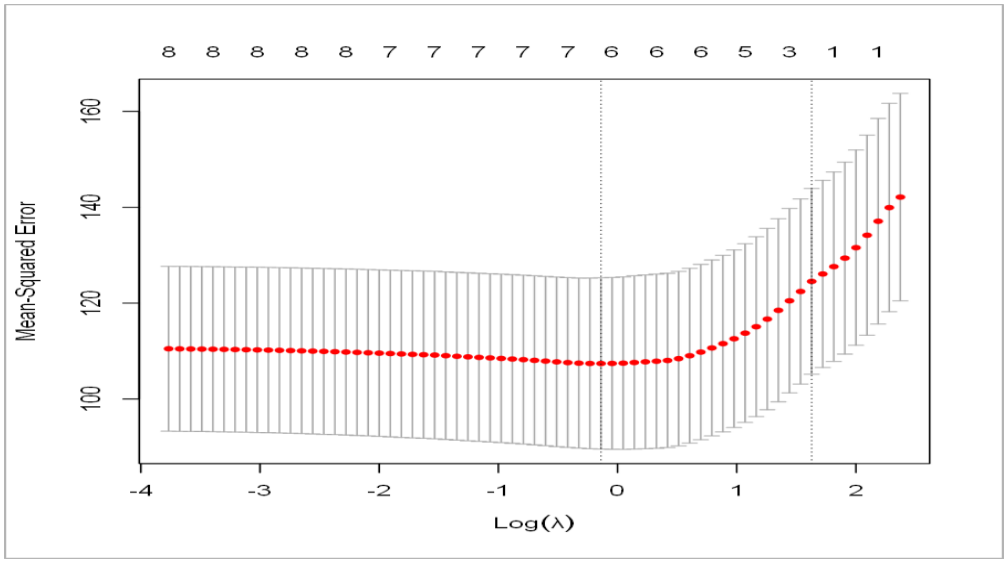
گروه تحلیل	گوسین
مقدار آلفا	۰/۵
تعداد مقادیر لامبدا	۱۰۰
استانداردسازی	انجام شده است
آستانه همگرایی	۰/۰۰۰۰۰۰۱
بیشینه از سرگیری	۱۰۰۰۰۰
تعداد بلوک برای اعتباریابی چندگانه	۱۰
داده‌های حاصل نشده	حذف شدند
مقدار لامبدای بهینه λ حاصل از اعتباریابی چندگانه	۰/۸۷

در این جدول مقدار آلفا همان‌گونه که گفته شد، برای رگرسیون الستیک شبکه‌ای ۰/۵ است و مقدار λ بهینه از طریق ۱۰ بلوک اعتباریابی برابر با ۰/۸۷ به‌دست آمده است.

جدول ۶- ضرایب خط رگرسیون الستیک

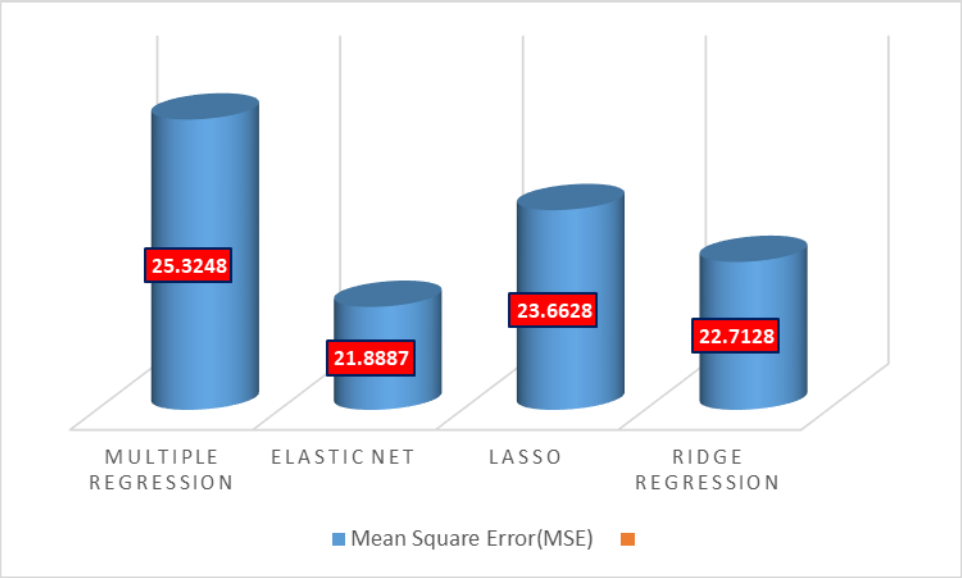
مقدارثابت	۴۵/۵۵
جنس	-۱۰/۰۶
وضعیت تاهل	-۰/۵۶
تحصیلات	صفر
شغل	صفر
سن	-۰/۱۷
تعدادفرزندان	۲/۸۱
سازش یافتگی	-۰/۳۱
اضطراب	۰/۸۲
متغیر وابسته: تجربه‌تروما	**R^2=۰/۶۲۱**

همان‌گونه که نتایج در جدول ۵ نشان می‌دهد، دو ضریب کوچک در جدول ۶، برابر با صفر شده است و میزان درصد واریانس تبیین شده افزایش یافته و به ۰/۶۲۱ رسیده است. نمودار ۱ اعتباریابی متقابل برای یافتن مقدار بهینه λ در رگرسیون الستیک شبکه‌ای را نشان می‌دهد. براساس این نمودار مقدار $\log(\lambda)$ حدود ۰/۷۲- نقطه بهینه برای شدت تابع جریمه (هزینه) مدل است. این نقطه جایی است که میزان MSE در اعتباریابی متقابل پس از آن افزایش یافته است.

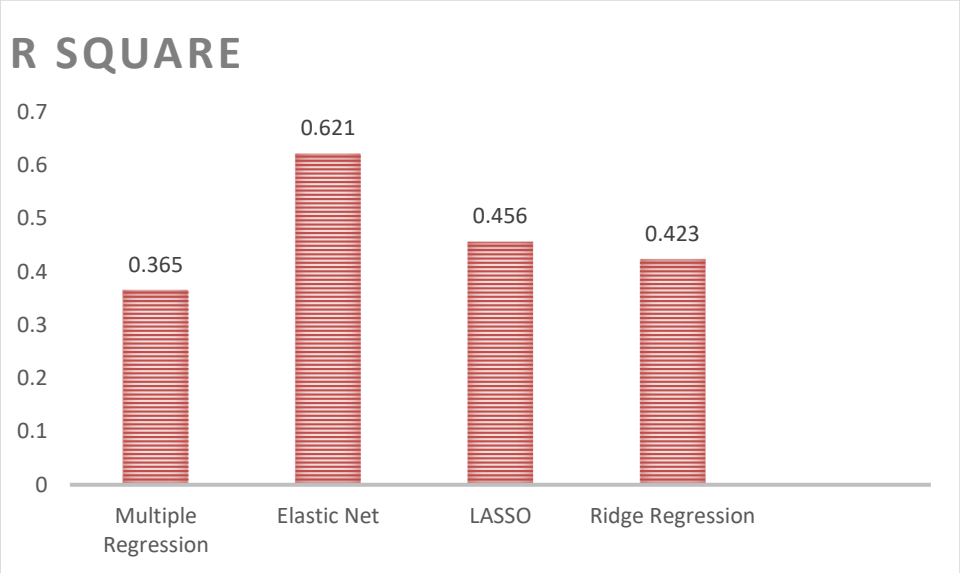


نمودار ۱- نمودار اعتبار متقابل برای یافتن مقدار بهینه در رگرسیون الستیک شبکه‌ای

برای مقایسه میزان دقت چهار مدل میانگین مجذور باقیمانده‌ها و مقادیر مجذور R موردمقایسه قرار گرفت. نمودارهای ۲ و ۳، مقادیر این شاخص را نشان می‌دهد.



نمودار ۲- مقادیر میانگین مجذور باقیمانده‌ها در چهار مدل رگرسیونی



نمودار ۳- مقادیر مجذور R در چهار مدل رگرسیونی

همان‌گونه نمودارهای ۲ و ۳ نشان می‌دهد، رگرسیون شبکه‌ای الستیک بهترین پیش‌بینی و رگرسیون چندگانه خطی کلاسیک ضعیف‌ترین پیش‌بینی را دارد.

بحث و نتیجه‌گیری

به‌طور خلاصه، در بسیاری از پژوهش‌ها در روان‌شناسی، تعداد متغیرها بسیار زیاد و تعداد مشاهدات (نمونه) اندک است (P>n)، که این، مشکلات زیادی را برای رگرسیون خطی ایجاد می‌کند. ازدیگرسو، وجود هم‌خطی چندگانه در متغیرهای پیش‌بین مشکل بزرگی است. در واقع، برخی از متغیرها ترکیب خطی از یک یا چند متغیر دیگر هستند؛ بنابراین در رویارویی با این مسائل می‌توان زیرمجموعه کوچکی از متغیرها را فراهم کرد که دارای بیشترین تأثیر باشند. ازسوی‌دیگر، توزیع نمرات باقیمانده در رگرسیون خطی چندگانه باید نرمال باشد. مفروضه‌ای کم‌تحقق یافته که نتیجه این تحلیل را در پژوهش‌های روان‌شناختی به خطری مهلک می‌کشاند دامگاهی ناپیدا برای کاربران این روش هستند، برای حل این مشکلات می‌توان از الگوریتم‌های انقباضی استفاده کرد. هدف الگوریتم‌های انقباضی (El net, Lasso, RR) برآورد ضرایب پیش‌بینی با ثبات در شرایطی است که متغیرهای پیش‌بینی همبستگی بالایی دارند. بسته به نوع جریمه‌ای که بر مدل وارد می‌شود با انواعی از الگوریتم‌های انقباضی روبه‌رو هستیم. در RR همه متغیرهای پیش‌بینی در مدل نگه داشته می‌شوند؛ در حالی‌که در Lasso ایجاد نتایج از طریق فشرده‌سازی برخی ضرایب به‌سمت صفر تضمین می‌شود. El net تلفیقی از RR و Lasso است. پیشنهاد می‌شود این روش‌ها با استفاده از داده‌های متعدد در زمینه‌های مختلف تکرار شود.

منابع

فراهانی، ح.، و عریضی، ح. (۱۳۸۷). *روش‌های پیشرفته پژوهش در علوم انسانی*. اصفهان: جهاد دانشگاهی اصفهان.

References

Cohen, J., Cohen, P., West, S., G., Aiken, L., S. (2003). *Applied multiple regression and correlation analysis for the behavioral sciences*. Third Edition. New York: Routledge.

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions. *PeerJ. 16(5)*, e3323.

Fox, J. (2016). *Applied regression analysis and generalized linear models (3rd Ed.)*. Thousand Oaks, CA: Sage publications.

Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2018). *Multivariate Data Analysis*. 8th edition: USA.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The Lasso and generalizations*. Chapman Hall, London.

Howell, D. (2013). *Statistical Methods for Psychology*. USA: Wadsworth.

Liu, H., & Zhang, J. (2009). Estimation consistency of the group Lasso and its applications. *J Mach Learn Res Workshop Conf Proc. 5*, 376–83.

Maronna, R. A. (2011). Robust ridge regression for high-dimensional data.

Technometrics. 53(1), 44-53.

Montgomery, D. C., Peck, E. A., &Vining, G. G. (2012). *Introduction to linear regression analysis*. USA: Wiley & Sons.

Park, H., & Konishi, S. (2015). Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *Journal of Statistical Computation and Simulation. 86(7),* 1-12.

Saleh, A. K. M. E., Arashi, M., & Kibria, B. M. G. (2019). *Theory of Ridge Regression Estimation with Applications*. Wiley, Hoboken, NJ, USA.

Samkar, H., & Alpu, O. (2010). Ridge regression based on some robust estimators. *Journal of Modern Applied Statistical Methods. 9(2).* 495-501.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics. 9(6),* 1135-1151.

Wilcox, R. R. (2019). Multicolinearity and ridge regression: results on type I errors, power and heteroscedasticity. *Journal of Applied Statistics. 46(5),* 946-957.

Williams, M. N., Grajales, C., & Kurkiewicz, D. (2013). Assumptions of multiple regression: correcting two misconceptions. Practical Assessment. *Research & Evaluation. 18(11),* 1–14.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society (B). 67(2),* 301-320.