Journal of Statistical Sciences, Autumn and Winter, 2023
Vol. 17, No. 2, pp 349-369
DOI: 10.52547/jss.17.2.07

Identification of Influential Observations for High–Dimensional Regression

Noori, N., Bevrani, H.
Department of Statistics, Faculty of Mathematics, Statistics and Computer Science, University of Tabriz, Tabriz, Iran.

Corresponding author: N. Noori, nasrin.nori95@gmail.com

Received: 26/6/2023 **Revised:** 21/9/2023 **Accepted and Published Online:** 24/9/2023.

Introduction

The prevalence of high-dimensional datasets has driven increased utilization of the penalized likelihood methods. However, when the number of observations is relatively few compared to the number of covariates, each observation can potentially tremendously influence model selection and inference. Therefore, identifying and assessing influential observations is vital in penalized methods. This article reviews measures of influence for detecting influential observations in high-dimensional lasso regression and has recently been introduced. Then, these measures under the elastic net method, which combines removing from lasso and reducing the ridge coefficients to improve the model predictions, are investigated. Through simulation and real datasets, illustrate that introduced influence measures effectively identify influential observations and can help reveal otherwise hidden relationships in the data.

Material and Methods

The deletion methods, like many linear regression diagnostics, measure the influence of an observation on lasso model selection by considering fitted lasso models when including versus excluding each of the observations and computing a scaled difference. The measures have been introduced are the: df-model a measure of the change in the model selected by the lasso; df-regpath a measure of the change in the lasso regularization path; df-cvpath a measure of the change in the lasso cross-validation path; and df-lambda a measure of the change in the lasso tuning parameter. As their names suggest, each measure targets a different aspect of lasso model selection. To ensure

our measures are easily applicable, theory is developed to demonstrate that these measures have tractable sampling distributions, allowing the use of external cut-offs for gauging influence. A novel aspect of this theory is that results are developed in the asymptotic framework even if $p \to \infty$: a setting that is most relevant for high-dimensional inference.

Results and Discussion

As observed from the simulation results, the influence measures under Elastic Net, similar to those under lasso, effectively identify influential observations. In addition, the influence measures under Elastic Net demonstrate good performance through actual data, as the results of identifying influential observations and essential genes after removing the identified influential observations by these measures and fitting the Elastic Net model, especially with $\alpha = 0.75$, do not significantly differ from the Lasso analysis output.

Conclusion

The need to identify influential observations is more critical in penalized regression than it is in ordinary least squares (OLS) regression. This is because in penalized regression, influential observations can, in addition to parameter estimates, also affect the model that is selected. We illustrated that in high-dimensional settings, individual observations could tremendously impact penalized model selection. Introducing measures could help reveal relationships in high-dimensional actual data that may otherwise remain hidden.

Keywords: Influence diagnostics, influential observations, high-dimensional data, penalized methods.

Mathematics Subject Classification (2010): 62J07, 62P10.

©The Author(s). The Publisher is Iranian Statistical Society.
This is an open access article distributed under the terms and conditions of (CC BY-NC 4.0)

مجله علوم آماری، پاییز و زمستان ۱۴۰۲
جلد ۱۷، شماره ۲، ص ۳۴۹ -- ۳۶۹
DOI: 10.52547/jss.17.2.07
مقاله پژوهشی

تشخیص مشاهدات موثر برای رگرسیون بعد بالا

نسرین نوری، حسین بیورانی
گروه آمار، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تبریز

نویسنده مسئول : نسرین نوری، nasrin.nori95@gmail.com

تاریخ دریافت: ۱۴۰۲/۴/۵ تاریخ بازنگری: ۱۴۰۲/۶/۳۰ تاریخ پذیرش و انتشار: ۱۴۰۲/۷/۲

چکیده: استفاده از روش‌های درستنمایی تاوانیده با تکثیر مجموعه داده‌های بعد بالا گسترش یافته است. با این وجود، هنگامی که تعداد مشاهدات در مقایسه با تعداد متغیرهای کمکی نسبتاً کم است، هر مشاهده‌ای به‌طور بالقوه می‌تواند تأثیر بسزایی روی انتخاب مدل و استنباط داشته باشد. بنابراین، شناسایی و ارزیابی مشاهدات موثر در روش‌های تاوانیده مهم است. در این مقاله، معیارهای تأثیر برای تشخیص مشاهدات موثر در رگرسیون لاسو بعد بالا که اخیراً معرفی شده‌اند، مرور می‌شوند. سپس، این معیارها تحت روش الاستیک‌نت که برای بهبود پیش‌بینی‌های مدل، ویژگی حذف از لاسو و کاهش ضرایب از ریج را ترکیب کرده، بررسی می‌شوند. از طریق شبیه‌سازی و مجموعه داده‌های واقعی نشان داده می‌شود که معیارهای تأثیر معرفی شده به‌طور کارآمد مشاهدات موثر را شناسایی می‌کنند و می‌توانند به آشکارسازی روابط پنهان در داده‌ها کمک کنند.

واژه‌های کلیدی: مباحث تشخیصی تأثیر، مشاهدات موثر، داده‌های بعد بالا، روش‌های تاوانیده.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62P10 ، 62J07.

۱ مقدمه

دسترسی به داده‌های با بعد بالا که در آن تعداد متغیرها (p) به طور قابل ملاحظه‌ای بیشتر از تعداد مشاهدات (n) است، اکنون در بسیاری از زمینه‌های علمی به ویژه ژنومیکس و زیست‌شناسی مولکولی رایج است. اغلب در تحلیل داده‌های بعد بالا فرض می‌شود تعداد متغیرهایی که در اصل با پاسخ موردنظر مرتبط هستند، کم است. بررسی در مورد

©نویسند(گان). ناشر انجمن آمار ایران است.
این مقاله با دسترسی آزاد تحت شرایط و ضوابط (CC BY-NC 4.0) توزیع شده است.

این تعداد کم از متغیرهای مهم به اهمیت روش‌های انتخاب مدل در محیط‌های بعد بالا تأکید کرده است. دسترسی گسترده به مجموعه داده‌های بعد بالا که $p >> n$، استفاده از روش‌های درستنمایی تاوانیده[1] را متداول ساخته است. با این حال، زمانی که تعداد مشاهدات نسبت به تعداد متغیرها کم است، هر مشاهده‌ای به‌طور بالقوه می‌تواند تأثیر چشمگیری روی انتخاب مدل و استنباط (برآورد پارامترها) داشته باشد. بنابراین، زمانی که از روش‌های تاوانیده برای داده‌های با بعد بالا استفاده می‌شود، تشخیص و ارزیابی مشاهدات موثر از اهمیت زیادی برخوردار است. در حقیقت، نیاز به پرداختن به این مسئله در رگرسیون تاوانیده نسبت به رگرسیون حداقل مربعات معمولی (OLS)[2] مهم‌تر است، چرا که مشاهدات موثر در رگرسیون تاوانیده می‌توانند علاوه بر برآورد پارامترها، روی مدلی که انتخاب می‌شود نیز تأثیر بگذارند.

با اینکه چن و همکاران (۲۰۱۰) و لامبرت و همکاران (۲۰۱۱) چندین روش لاسو استوار[3] را پیشنهاد داده‌اند و شی و همکاران (۲۰۱۱) با استفاده از رگرسیون تاوانیده غیرمحدب[4]، تشخیص داده دورافتاده[5] را بررسی کرده‌اند، مطالعات اندکی شناسایی مشاهدات موثر در رگرسیون لاسو را بررسی کرده‌اند. ژائو و همکاران (۲۰۱۳) شیوه‌ای در این راستا برای بستر $p > n$ پیشنهاد کردند. از آنجایی که برآورد رگرسیون OLS به‌طور منحصربه‌فرد در این محیط تعریف نمی‌شود، ژائو و همکاران (۲۰۱۳) در عوض اندازه‌گیری تأثیر یک مشاهده را با استفاده از همبستگی‌های حاشیه‌ای[6] پیشنهاد کردند. وانگ و لی (۲۰۱۷) روشی برای تشخیص نقطه دورافتاده در مقابل تشخیص مشاهدات موثر در مدل‌های رگرسیون بعد بالا پیشنهاد کردند. روش وانگ و لی (۲۰۱۷) بر اساس تفاوت‌هایی در همبستگی فاصله است. برای شناسایی مشاهدات موثر در داده‌های با بعد بالا می‌توان چارچوب ارزیابی تأثیر برآورد در رگرسیون خطی کلاسیک را توسعه و نشان داد که به همان اندازه (اگر نه بیشتر) برای ارزیابی تأثیر انتخاب مدل در رگرسیون بعد بالا مرتبط است. از اواخر دهه ۱۹۷۰ تحقیقات راهگشا در مباحث تشخیصی تأثیر برای رگرسیون OLS شروع شد. برای مثال از بلزلی و همکاران (۱۹۸۰)، کوک و ویزبرگ (۱۹۸۲)، آتکینسون (۱۹۸۱) و کوک (۱۹۷۷، ۱۹۷۹) می‌توان نام برد. به ویژه، بلزلی و همکاران (۱۹۸۰) چارچوبی برای رگرسیون خطی کلاسیک جهت تشخیص مشاهدات موثر پیشنهاد کردند که امروزه هنگام مشاهده مشکلات تأثیر در رگرسیون بعد بالا به همان اندازه مرتبط است. آتکینسون (۱۹۸۴) مرور روشنگرانه‌ای از بلزلی و همکاران (۱۹۸۰) و کوک و ویزبرگ (۱۹۸۲) و بررسی در زمینه مباحث تشخیصی رگرسیون در آن زمان ارائه کردند. برخی از پرکاربردترین معیارها، تأثیر یک مشاهده را با مقایسه جنبه‌های خاصی از مدل برازش شده در حضور و عدم حضور مشاهده موردبحث می‌سنجند. مثال‌هایی از برخی از فراگیرترین این معیارها عبارتند: Cook's D و DFBETA که تغییرات در برآوردهای ضرایب و DFFITS که تغییرات در مقادیر برازش شده را می‌سنجند. نکته مهم این است که مباحث تشخیصی تأثیر کلاسیک بر تأثیر روی

[1] Penalized likelihood methods
[2] Ordinary Least Squares
[3] Robust lasso procedures
[4] Nonconvex penalized regression
[5] Outlier
[6] Marginal correlations

برآورد پارامتر تمرکز می‌کند به این معنی است که معیارهای جدیدی در زمینه رگرسیون بعد بالا مورد نیاز هستند که تأثیر بر انتخاب مدل را هدف قرار دهند.

در بخش ۲، معیارهای تأثیر پیشنهاد شده برای رگرسیون لاسو توسط راجاراتنام و همکاران (۲۰۱۹) مرور می‌شوند. در این بخش، چهار روش حذفی با توسعه چارچوب ارزیابی تأثیر برآورد در رگرسیون خطی کلاسیک برای سنجش تأثیر یک مشاهده روی انتخاب مدل لاسو معرفی می‌شوند. سپس به معرفی رگرسیون الاستیک‌نت در بخش ۳ پرداخته می‌شود. در بخش ۴، مطالعه شبیه‌سازی به منظور بررسی و مقایسه عملکرد معیارهای تأثیر لاسو با همان معیارها تحت الاستیک‌نت به‌ازای αهای مختلف ارائه می‌شود. داده‌های مربوط به توصیف ژن گلیوبلاستما[1] را در بخش ۵ معرفی کرده و مباحث نظری روی این داده‌ها پیاده می‌شود (نوری ، ۱۴۰۱). در نهایت، بحث و نتیجه‌گیری ارائه می‌شود.

۲ معیارهای تأثیر برای رگرسیون لاسو

لاسو تعمیم‌یافته رگرسیون OLS است که توسط تیبشیرانی (۱۹۹۶) پیشنهاد شد. لاسو با تحمیل جریمه ℓ_1 یک مدل تنک[2] تولید می‌کند که تعدادی از برآوردهای ضرایب را دقیقاً برابر صفر قرار می‌دهد. برآوردهای ضرایب رگسیونی لاسو را می‌توان به صورت

$$\hat{\underset{\sim}{\beta}} = \arg\min_{\underset{\sim}{\beta}} \Big\{ \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 \Big\} + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (1)$$

تعریف کرد که $\hat{\underset{\sim}{\beta}}$ یک بردار pـبُعدی شامل برآوردهای ضرایب لاسو و λ یک پارامتر منظم‌سازی[3] است. در عمل می‌توان فرض کرد، متغیر پاسخ میانگین ۰ دارد و متغیرهای کمکی استاندارد شده دارای میانگین ۰ و واریانس ۱ هستند. جریمه ℓ_1 روی ضرایب این خاصیت را دارد که با افزایش λ، ضرایب به سمت صفر منقبض شده و تعدادی از ضرایب نیز دقیقاً به مقدار ۰ کاهش می‌یابند، اگر λ به اندازه کافی بزرگ باشد. این توانایی به لاسو اجازه اجرای انتخاب مدل و تولید نتایج قابل تفسیر را می‌دهد. نتایج تحلیل لاسو که در آن پارامتر منظم‌سازی با اعتبارسنجی متقابل[4] برآورد شده، می‌تواند توسط چهار خروجی مربوطه خلاصه شود: الف- مدل انتخابی به وسیله لاسو و برآوردهای ضرایب مربوطه؛ ب- مسیر منظم‌سازی؛ ج- پارامتر منظم‌سازی برآورد شده $(\hat{\lambda})$؛ د- خطای اعتبارسنجی متقابل به عنوان تابعی از λ. در حقیقت این چهار کمیت به ویژه مدل انتخاب شده توسط لاسو، جزء جدایی‌ناپذیر راه‌حل لاسو هستند که نظارت بر تغییرات این مقادیر می‌تواند راهنمای مفیدی برای تشخیص تأثیر باشد.

[1] Glioblastoma Gene Expression Data
[2] Sparse Model
[3] Regularization parameter
[4] Cross-validation

۲.۱ معیار Df-Model

اوّلین معیار تأثیری که معرفی می‌شود، df-model است که به طور مستقیم تغییر در مدل انتخاب شده توسط لاسو را زمانی که مشاهده‌ای حذف شده، ارزیابی می‌کند. تعیین اندازه این تغییر مهم است چون یک تغییر بزرگ در مدل انتخاب شده توسط لاسو می‌تواند به طور چشمگیری نتایج حاصل از تجزیه و تحلیل را تغییر دهد. Df-model برای iـامین مشاهده (راجاراتنام و همکاران ، ۲۰۱۹) به صورت

$$df - \mathrm{model}(i) = \frac{\delta(i) - \mathrm{E}\{\delta(i)\}}{\sqrt{\mathrm{Var}\{\delta(i)\}}} \tag{۲}$$

تعریف شده است که $\delta(i) = \sum_{j=1}^{p} \left| I\{\hat{\beta}_j^{\mathrm{lasso}} = 0\} - I\{\hat{\beta}_j^{\mathrm{lasso}}(i) = 0\} \right|$ و $I\{\cdot\}$ تابع نشانگر را نشان می‌دهد. به ویژه، df-model یک معیار مقیاس شده از تعداد تغییرات در متغیرهای پیشگوی انتخاب شده است که در راه‌حل لاسو به هنگام حذف یک مشاهده روی می‌دهد.

محاسبه df-model شامل برازش $n+1$ بار لاسو است و پس از آن مقادیر مشاهده شده $\delta(i)$ حساب می‌شوند. میانگین نمونه و واریانس n مقدار مشاهده شده $\delta(i)$ می‌توانند به ترتیب به عنوان برآوردهای $\mathrm{E}\{\delta(i)\}$ و $\mathrm{Var}\{\delta(i)\}$ به کار روند. نظریه‌ای که در ادامه اثبات می‌کند، مقادیر محاسبه شده df-model را می‌توان با مقادیر بُرینش ± 2 مقایسه کرد.

نتیجه مقادیر بُرینش برای df-model در محیط طرح متعامد[1] بعد بالا نشان داده خواهد شد. این نتیجه مقادیر بُرینش طبیعی و قدرمطلق مشاهداتی با بزرگی df-model(i) بیش از ۲ را می‌دهد که می‌توان برای بررسی بیشتر نشانه‌گذاری کرد.

قضیه ۱. (راجاراتنام و همکاران ، ۲۰۱۹) مدل رگرسیونی تعریف شده به صورت

$$y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$$

برای هر $i \in \{1, \ldots, n\}$ در نظر بگیرید که $X'X = I_p$ (یعنی ماتریس طرح متعامد است)، $n \geq p$ و p اجازه رشد بدون محدودیت را دارد؛ متغیر پاسخ در مرکز قرار گرفته است تا میانگین ۰ داشته باشد؛ پیشگوهای X_{ij} استاندارد شده‌اند تا میانگین ۰ و واریانس ۱ داشته باشند و خطاهای ϵ_i متغیرهای تصادفی مستقل با توزیع $\mathrm{N}(0, \sigma^2)$ که $\sigma^2 > 0$ مجهول است. حال فرض می‌شود

$$\Delta = \sum_{j=1}^{p} \delta_j, \qquad \delta(j) = \left| I\{\beta_j^{\mathrm{true}} = 0\} - I\{\hat{\beta}_j^{\mathrm{lasso}} = 0\} \right|, \qquad j \in \{1, \ldots, p\}.$$

[1] Orthogonal design setting

همچنین فرض می‌شود $\hat{\beta}_j^{\text{lasso}}(s)$، برآوردگر لاسو برای ضریب jـام را نشان می‌دهد زمانی که پارامتر جریمه ℓ_1 در درستنمایی مقدار s می‌گیرد. اگر دنباله‌ای از مقادیر ضرایب واقعی صدق کند در

$$\sum_{j=1}^{p} I\{\beta_j^{\text{true}} = 0\} \to \infty,$$

آنگاه $p \to \infty$ (با $n \equiv n_p \to \infty$). بنابراین دنباله‌ای از متغیرهای تصادفی $\Delta \equiv \Delta_p$ صدق می‌کند در

$$\frac{\Delta_p - \mathrm{E}(\Delta_p)}{\sqrt{\mathrm{Var}(\Delta_p)}} \rightsquigarrow \mathrm{N}(0, 1).$$

۲.۲ معیار Df-Regpath

حال معیار تأثیر df-regpath معرفی می‌شود تا تغییر کلی در مسیر منظم‌سازی راه‌حل لاسو هنگامی که یک مشاهده معین حذف شده، سنجیده شود. ارزیابی این تغییر مهم است زیرا یک تغییر بزرگ در مسیر منظم‌سازی به این معنی است که نمایه ضرایب برآورد شده لاسو به دلیل حذف یک مشاهده به طور قابل توجهی تغییر کرده است. چنین تغییراتی می‌تواند به این معنی باشد که مشاهده موردنظر تأثیر قابل ملاحظه‌ای روی ضرایب برآورد شده لاسو و به طور بالقوه تغییر در تفسیر و نتایج حاصل از تجزیه و تحلیل لاسو دارد. Df-regpath برای iـامین مشاهده به صورت (راجاراتنام و همکاران ، ۲۰۱۹)

$$df - \text{regpath}(i) = \frac{\Delta_1 \hat{\underset{\sim}{\beta}}^{\text{lasso}}(i) - \mathrm{E}\{\Delta_1 \hat{\underset{\sim}{\beta}}^{\text{lasso}}(i)\}}{\sqrt{\mathrm{Var}\{\Delta_1 \hat{\underset{\sim}{\beta}}^{\text{lasso}}(i)\}}} \tag{۳}$$

تعریف شده است، که در آن $\Delta_1 \hat{\underset{\sim}{\beta}}^{\text{lasso}}(i) = \int_l^u \left\| \hat{\underset{\sim}{\beta}}^{\text{lasso}}(s) - \hat{\underset{\sim}{\beta}}^{\text{lasso}}(s, i) \right\|_1 ds$ و l و u به گونه‌ای مشخص شده‌اند که بازه $[l, u]$ محدوده‌ای از مقادیر قابل قبول λ را تعریف می‌کند. به طور خاص، df-regpath یک معیار مقیاس شده از تغییر کلی در مسیر منظم‌سازی لاسو به هنگام حذف یک مشاهده است.

محاسبه df-regpath شامل برازش $n + 1$ بار لاسو روی یک دنباله از مقادیر λ است که محدوده $[l, u]$ را در بر می‌گیرد. از این $n + 1$ مدل لاسو برازش شده، نمایه‌های $\hat{\underset{\sim}{\beta}}^{\text{lasso}}(s)$ و $\hat{\underset{\sim}{\beta}}^{\text{lasso}}(s, i)$ به دست می‌آیند. سپس انتگرال $\Delta_1 \hat{\underset{\sim}{\beta}}^{\text{lasso}}(i)$ با استفاده از تکنیک‌های عددی استاندارد که برای مقادیر $\left\| \hat{\underset{\sim}{\beta}}^{\text{lasso}}(s) - \hat{\underset{\sim}{\beta}}^{\text{lasso}}(s, i) \right\|_1$ روی دنباله‌ای از مقادیر λ اعمال شده، تقریب می‌یابد. میانگین نمونه و واریانس n مقدار مشاهده شده $\Delta_1 \hat{\underset{\sim}{\beta}}^{\text{lasso}}(i)$ را می‌توان به ترتیب به عنوان برآوردهای $\mathrm{E}\{\Delta_1 \hat{\underset{\sim}{\beta}}^{\text{lasso}}(i)\}$ و $\mathrm{Var}\{\Delta_1 \hat{\underset{\sim}{\beta}}^{\text{lasso}}(i)\}$ استفاده کرد. دوباره مشابه df-model، نظریه‌ای که در ادامه نشان می‌دهد می‌توان مقادیر محاسبه شده df-regpath را با مقادیر بُرینش ± 2 مقایسه کرد.

در این بخش، تجزیه و تحلیل دقیقی از معیار تأثیر df-regpath در محیط طراحی متعامد انجام خواهد شد. بنابراین، اکنون در مورد یک قضیه بحث می‌شود که تغییرپذیری نمونه‌گیری معیار تأثیر df-regpath را در محیط طراحی متعامد اندازه می‌گیرد.

قضیه ۲. مدل رگرسیونی تنک توصیف شده در قضیه ۱ را در نظر بگیرید و فرض کنید وجود دارد $\eta < ۱$، $\epsilon > ۰$ و $M > ۰$ به طوری که

$$\lim_{p\to\infty} \sup \left[\frac{۱}{p}\sum_{j=۱}^{p} I\{\beta_j^{\text{true}} \neq ۰\}\right] = \eta$$

و برای همه $p \geq ۱$،

$$\frac{۱}{p}\sum_{j=۱}^{p} |\beta_j^{\text{true}}|^{۴+\epsilon} \leq M. \tag{۴}$$

آنگاه:

الف- امید ریاضی $\mu_j(\lambda) = \mathrm{E}[\hat{\beta}_j^{\text{lasso}}(\lambda)]$ برای هر $j \in \{۱, \ldots, p\}$ و برای همه $\lambda > ۰$ متناهی است؛

ب- انتگرال تصادفی

$$\Omega_p = \int_۰^\infty \left\|\hat{\underset{\sim}{\beta}}^{\text{lasso}}(\lambda) - \underset{\sim}{\mu}(\lambda)\right\|_۱ d\lambda = \int_۰^\infty \sum_{j=۱}^{p} \left|\hat{\beta}_j^{\text{lasso}}(\lambda) - \mu_j(\lambda)\right| d\lambda$$

برای همه بردارهای پاسخ ممکن $\underset{\sim}{y} \in \mathbb{R}^n$ متناهی است (یعنی Ω_p یک متغیر تصادفی خوب تعریف شده است) که $\underset{\sim}{\mu}(\lambda)$ برداری با طول p و با j-امین عنصر $\mu_j(\lambda)$ است؛

ج- $۰ < \mathrm{E}(\Omega_p) < \infty$ و $۰ < \mathrm{Var}(\Omega_p) < \infty$؛

د- وقتی $p \to \infty$، $\frac{\Omega_p - \mathrm{E}(\Omega_p)}{\sqrt{\mathrm{Var}(\Omega_p)}} \rightsquigarrow \mathrm{N}(۰, ۱)$.

برهان: اثبات قضیه ۲ شامل متناهی کردن امید و واریانس انتگرال‌های تصادفی بالا و استفاده از قضیه حد مرکزی لیاپانوف برای قسمت (ت) است (راجاراتنام و همکاران ، ۲۰۱۹).

۲.۳ معیار Df-Cvpath

اکنون معیار تأثیر df-cvpath معرفی می‌شود تا تغییر در عملکرد پیشگویی لاسو هنگامی که مشاهده‌ای حذف شده، ارزیابی شود. محاسبه این تغییر مهم است چرا که تغییر بزرگی در عملکرد پیشگویی لاسو بسیار تذکردهنده خواهد بود که مشاهده موردنظر تأثیر قابل توجهی بر راه‌حل لاسو و یک مقدار پاسخ غیرمعمول دارد. منحنی خطای اعتبارسنجی متقابل $\gamma(s)$ برای مقادیر متفاوت پارامتر منظم‌سازی $\lambda = s$، خطای پیشگویی روی داده‌های آزمون را بعد از اینکه

روش لاسو روی مولفه متفاوت داده‌ها آموزش دیده، اندازه می‌گیرد. Df-cvpath برای iـامین مشاهده به صورت (راجاراتنام و همکاران ، ۲۰۱۹)

$$df-\text{cvpath}(i) = \frac{\Delta\gamma(i) - \text{E}\{\Delta\gamma(i)\}}{\sqrt{\text{Var}\{\Delta\gamma(i)\}}} \tag{۵}$$

تعریف شده است که در آن $\Delta\gamma(i) = \int_l^u |\gamma(s) - \gamma(s,i)| ds$ و $\gamma(s,i)$ خطای اعتبارسنجی متقابل هنگامی که iـامین مشاهده حذف شده، است. l و u به گونه‌ای تعریف شده‌اند که بازه $[l,u]$ محدوده‌ای از مقادیر قابل قبول λ را معین می‌کند. به طور مشخص، df-cvpath یک معیار مقیاس شده از تغییر کلی در مسیر اعتبارسنجی متقابل لاسو زمانی که مشاهده‌ای حذف شده، است.

Df-cvpath را می‌توان با روشی مشابه df-regpath محاسبه کرد. یک رویکرد مشابه با روشی که جهت ایجاد مقادیر بُرینش برای df-regpath استفاده شده را می‌توان برای توجیه مقادیر بُرینش تقریبی ۲± برای df-cvpath به کار برد. برای رعایت اختصار از جزئیات صرف‌نظر شده است.

۲.۴ معیار Df-Lambda

حال معیار تأثیر df-lambda معرفی می‌شود تا تغییر در مقدار بهینه پارامتر منظم‌سازی لاسو λ زمانی که مشاهده‌ای حذف شده، سنجیده شود. محاسبه این تغییر مهم است زیرا تغییر بزرگی در مقدار بهینه λ نشان می‌دهد که مشاهده موردنظر تأثیر زیادی بر میزان کاهش برآوردهای ضرایب مدل انتخابی لاسو دارد. Df-lambda برای iـامین مشاهده به صورت (راجاراتنام و همکاران ، ۲۰۱۹)

$$df-\text{lambda}(i) = \frac{\hat{\lambda} - \hat{\lambda}(i) - \text{E}\{\hat{\lambda} - \hat{\lambda}(i)\}}{\sqrt{\text{Var}\{\hat{\lambda} - \hat{\lambda}(i)\}}} \tag{۶}$$

تعریف شده است. به ویژه، df-lambda یک معیار مقیاس شده از تفاوت بین مقدار بهینه λ بر اساس کل مجموعه داده ($\hat{\lambda}$) و مقدار بهینه λ زمانی که iـامین مشاهده حذف شده ($\hat{\lambda}(i)$)، است.

محاسبه df-lambda شامل برازش $n+1$ بار لاسو است تا مقادیر $\hat{\lambda} - \hat{\lambda}(i)$ به دست آیند. میانگین نمونه و واریانس n مقدار مشاهده شده $\hat{\lambda} - \hat{\lambda}(i)$ را می‌توان به ترتیب به عنوان برآوردهای $\text{E}\{\hat{\lambda} - \hat{\lambda}(i)\}$ و $\text{Var}\{\hat{\lambda} - \hat{\lambda}(i)\}$ استفاده کرد.

مقادیر بُرینش ۲± برای df-lambda را می‌توان از نظر اکتشافی توجیه کرد. برای kهای بزرگ که k تعداد دسته‌ها در اعتبارسنجی متقابل است، خطای اعتبارسنجی متقابل به صورت نرمال توزیع شده است. این امر با

تشخیص خطای اعتبارسنجی متقابل $\gamma(\hat{\lambda}) = \frac{1}{k}\sum_{i=1}^{k} PE_i(\hat{\lambda})$ (که $PE_i(\hat{\lambda})$ خطای پیشگویی روی iـامین دسته از داده‌ها را نشان می‌دهد) که یک میانگین است، دنبال می‌شود و پس از آن می‌توان قضیه حد مرکزی را اعمال کرد. فراخوانی روش دلتا[1] و معکوس کردن $\gamma(\hat{\lambda})$ تا $\hat{\lambda}$ به دست آید. همچنین نتیجه گرفته می‌شود که $\hat{\lambda}$ تقریباً به صورت نرمال با میانگین و واریانس متناظر توزیع شده است. بنابراین، کمیت استاندارد شده در معادله (۶) را می‌توان با مقادیر بُرینش تقریبی ± 2 استفاده کرد.

۳ رگرسیون الاستیک‌نت

از محدودیت‌های لاسو می‌توان به موارد زیر اشاره کرد،

الف- اگر $p > n$ باشد، لاسو حداکثر n متغیر می‌تواند انتخاب کند، در حالی که این احتمال وجود دارد، بیش از n متغیر یا اینکه همه متغیرها (p) مرتبط با متغیر پاسخ باشند.

ب- اگر مجموعه‌ای از متغیرها رو داشته باشیم که دارای هم‌خطی بالایی باشند، لاسو فقط یک متغیر آن هم به صورت تصادفی انتخاب کرده و از بقیه متغیرهای مجموعه چشم‌پوشی می‌کند که این برای تکرارپذیری و تفسیر داده‌ها خوب نیست.

با توجه به محدودیت های لاسو، رگرسیون الاستیک‌نت که ترکیبی از رگرسیون ریج[2] (هورل و کنارد ، ۱۹۷۰) و لاسو (تیبشیرانی ، ۱۹۹۶) است، توسط زو و هستی (۲۰۰۵) معرفی شد. آنها این روش را برای تعدیل همزمان مشکلات همخطی چندگانه و تفسیرناپذیر بودن مدل مطرح کردند (معنوی و روزبه ، ۱۳۹۹). مسئله بهینه‌سازی به صورت

$$\min_{\beta\in\mathbb{R}^p}\Big\{\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\Big\} \tag{۷}$$

تعریف می‌شود. رگرسیون الاستیک‌نت دارای دو پارامتر λ_1 و λ_2 است که طبیعتاً یافتن مقادیر مناسب برای دو پارامتر تاوان در این روش نسبت به روش‌هایی که تنها یک پارامتر تاوان دارند، امری به مراتب دشوارتر خواهد بود و این یکی از معایب این روش محسوب می‌شود. با در نظر گرفتن $\alpha = \lambda_1/(\lambda_1+\lambda_2)$ می‌توان برآوردگر الاستیک‌نت را به فرم تاوانی

$$\hat{\beta} = \min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \qquad s.t. \qquad \alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1 \le t \tag{۸}$$

[1] Delta method
[2] Ridge regression

نوشت. ناحیه تاوان در این روش ترکیبی، اکیداً محدب از نواحی تاوان در روش‌های ریج و لاسو است. این فرم تابع تاوان، علاوه بر قابلیت صفر برآورد کردن برخی از ضرایب، توانایی برابر برآورد کردن ضرایب متغیرهایی که اثر یکسان روی متغیر پاسخ دارند یا به شدت همبسته هستند را نیز دارد. به این ترتیب از این روش می‌توان برای گروه‌بندی متغیرهای پیشگو استفاده کرد به خصوص زمانی که تعداد آنها زیاد است. برآوردگر اصلاح شده الاستیک‌نت که دارای دقت پیش‌بینی بالاتری نسبت به برآوردگر اوّلیه است به صورت $\hat{\beta}^{ENet} = (۱ + \lambda_۲)\hat{\beta}$ تعریف می‌شود که در آن $\hat{\beta}^{ENet}$ و $\hat{\beta}$ به ترتیب برآوردگرهای الاستیک‌نت و الاستیک‌نت اوّلیه (خام) است.

۴ مطالعه شبیه‌سازی

یک مجموعه ایده‌آل با $n = ۵۰$ و $p = ۱۰۰۰$ را در نظر بگیرید که پنج متغیر کمکی اوّل از طریق

$$Y = ۱X_۱ + ۲X_۲ + ۳X_۳ + ۴X_۴ + ۵X_۵ + \epsilon \qquad (۹)$$

با متغیر پاسخ مرتبط هستند که ϵ یک نمونه تصادفی با توزیع $\mathrm{N}(۰, ۱)$ است؛ ماتریس متغیر کمکی ۵۰×۱۰۰۰ به صورت نرمال چندمتغیره با میانگین ۰ و واریانس ۱ تولید می‌شود و همبستگی دوبه‌دو بین X_i و X_j برابر است با $\rho^{|i-j|}$. این داده‌های نرمال شبیه‌سازی شده مربوط به وضعیتی است که در آن تمام مفروضات رگرسیون OLS استاندارد برقرار است و می‌توان انتظار داشت که حذف مشاهدات فردی، کمترین تأثیر را بر روی راه‌حل لاسو داشته باشد. حال به شبیه‌سازی‌ها پرداخته می‌شود تا توانایی معیارهای معرفی شده جهت شناسایی مشاهدات موثر ارزیابی شود. ابتدا مهم است تا بهتر شناخته شوند انواع مشاهداتی که می‌توانند منجر به تغییراتی در مدل انتخاب شده توسط لاسو گردند. علاوه بر آن ضروری است تا مشخص شود نوع مشاهداتی که می‌توانند منجر به بی‌ثباتی مدل یا تغییراتی در تنکی راه‌حل لاسو شوند. این مشخص‌سازی با قرارگیری مشاهدات بالقوه موثر در مثال‌های شبیه‌سازی زیر کمک می‌کند. مراحل زیر جهت تولید داده‌ها برای مطالعه شبیه‌سازی دنبال شدند:

۱- ماتریس متغیرهای کمکی ۵۰×۱۰۰۰ از توزیع نرمال چندمتغیره با میانگین ۰ و واریانس ۱ تولید کنید و همبستگی دوبه‌دو بین X_i و X_j برابر است با $\rho^{|i-j|}$. هر دو طرح متعامد ($\rho = ۰$) و نامتعامد ($\rho = ۰/۵$) در نظر گرفته خواهد شد.

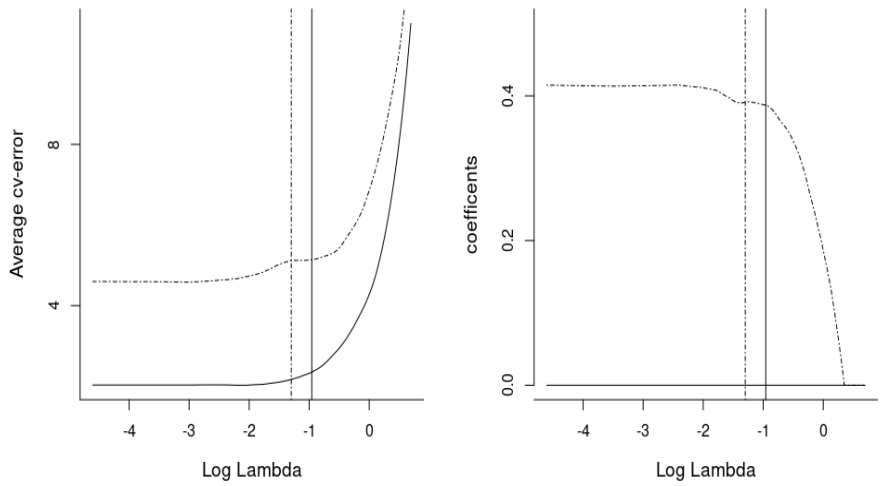
۲- از مرحله ۱، ۱۰۰ـامین عنصر مشاهده ۱ ($x_{۱,۱۰۰}$) را با مقدار a جایگزین کنید.

۳- مقادیر پاسخ را با توجه به رابطه تعریف شده توسط معادله (۹) شبیه‌سازی کنید.

۴- از مرحله ۳، مقدار پاسخ شبیه‌سازی شده برای مشاهده ۱ ($y_۱$) را با $\mu(y_۱) + b$ جایگزین کنید که در آن $\mu(y_۱)$ مقدار میانگین $y_۱$ داده شده توسط معادله (۹) است.

تنظیم $|a|$ یا $|b|$ به اندازه کافی بزرگ اجازه می‌دهد یک مجموعه داده تولید شود که مشاهده ۱ متغیر کمکی یا مقدار(های) پاسخ بزرگ (در قدر مطلق) غیرعادی دارد. این نوع مشاهده انتظار می‌رود که تأثیر زیادی بر انتخاب

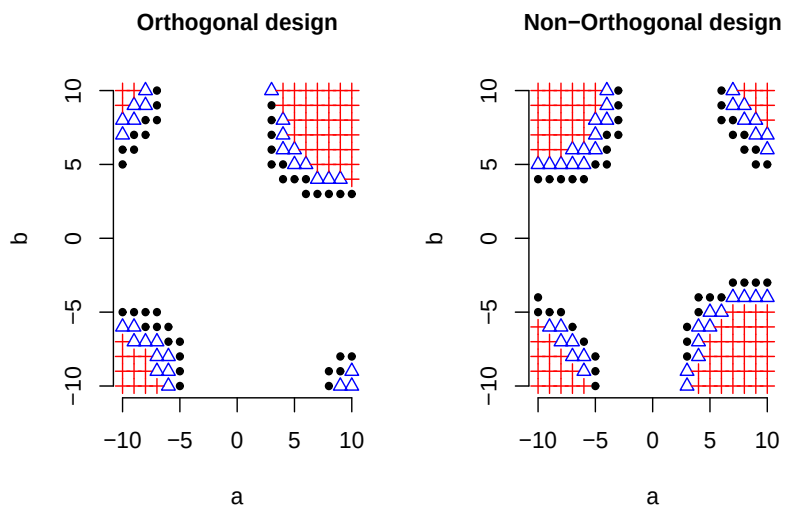
مدل لاسو داشته باشد. به طور مشخص از آنجایی که مقدار $x_{۱,۱۰۰}$ را مختل می‌شود، انتظار می‌رود (ممکن است) که تغییری در تنکی برای $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ در میان اثرات دیگر به ازای $|a|$ و $|b|$ بزرگ مشاهده شود.



شکل ۱. نمودار خطای اعتبارسنجی متقابل (چپ) و مسیر منظم‌سازی (راست) $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ از لاسو برازش شده به داده‌های طراحی نامتعامد شبیه‌سازی شده با $a = b = ۱۰$ در حضور و عدم حضور مشاهده ۱

به عنوان مقدمه‌ای برای شبیه‌سازی‌ها نشان داده خواهد شد که چگونه یک مشاهده موثر منحصربه‌فرد می‌تواند بر جنبه‌های خاصی از راه‌حل لاسو تأثیر بگذارد. یک مجموعه داده طراحی نامتعامد با $a = b = ۱۰$ تولید و بررسی می‌شود که چگونه خطای اعتبارسنجی متقابل و مسیر منظم‌سازی هنگامی که مشاهده موثر (مشاهده ۱) حذف شده، تغییر می‌کنند. شکل ۱ شامل نمودار خطای اعتبارسنجی متقابل و مسیر منظم‌سازی برای $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ از لاسو برازش شده به مجموعه داده شبیه‌سازی شده با و بدون مشاهده ۱ در مدل می‌شود. در این شکل، خطوط توپر مربوط به لاسو برازش شده هنگام حذف مشاهده ۱ است و خطوط عمودی با مقدار بهینه λ مطابقت دارند. به وضوح، حذف مشاهده ۱ افت قابل ملاحظه‌ای در خطاهای اعتبارسنجی متقابل و افزایشی در مقدار بهینه λ را نتیجه می‌دهد. مسیر منظم‌سازی $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ برای لاسو کامل در مقابل لاسو برازش شده بدون مشاهده ۱ نیز به طور اساسی متفاوت است. این مسیر برای $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ زمانی که مشاهده ۱ حذف شده در صفر صاف است یعنی پیشگوی $X_{۱۰۰}$ هرگز وارد مدل نمی‌شود. با این حال هنگامی که مشاهده ۱ شامل شود، مسیر برای $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ تا زمانی که $\log(\lambda)$ نزدیک صفر نباشد به صفر نمی‌رسد. از شکل ۱ مشخص است که $\hat{\beta}^{\text{lasso}}_{۱۰۰} > ۰$ و $\hat{\beta}^{\text{lasso}}_{۱۰۰}(۱) = ۰$. این مثال نشان می‌دهد که مشاهدات موثر می‌توانند بر خطای اعتبارسنجی متقابل، مسیر منظم‌سازی، مقدار بهینه λ و مدل انتخاب شده تأثیر بگذارند. در حقیقت، یک نقطه موثر می‌تواند به طور بالقوه بر همه چهار معیار تأثیر بگذارد.

مهم است برای طراحی شبیه‌سازی‌ها مقادیر a و b پیدا شوند که احتمالاً منجر به تغییر در تنکی برای $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ شدند. برای بررسی این سوال، مجموعه داده‌های طراحی متعامد و نامتعامد تولید شدند. سپس مقادیر a و b بررسی شدند که منتهی به تغییر در تنکی برای $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ هنگام حذف مشاهده ۱ شدند. به ویژه با استفاده از مرحله ۱ طرح تولید داده، داده‌های متغیر کمکی تولید شدند. هر بار با توجه به اینکه آیا حذف مشاهده ۱ سبب تغییر در تنکی برای $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ شد، مراحل ۲-۴ با استفاده از داده‌های تولید شده روی شبکه‌ای از مقادیر a و b تکرار شدند. این فرایند منتهی شد به محاسبه درصد دفعاتی که یک ترکیب خاص از a و b منجر به تغییر در تنکی شد. نتایج این تحلیل در شکل ۲ گزارش می‌شود. به بیان دقیق‌تر برای هر جفت مقادیر a و b روی یک شبکه، نسبت دفعاتی که $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ تغییر در تُنُکی دارد به تصویر کشیده می‌شود. نمادهای $+$، $\triangle$ و $\bullet$ به ترتیب مربوط به $p > ۰/۷۵$، $۰/۵ < p \leq ۰/۷۵$ و $۰/۲۵ < p \leq ۰/۵$ هستند. این شکل‌ها نشان می‌دهند که تغییرات در تنکی در نواحی رخ می‌دهند که تقریباً با $|a| \geq ۵$ و $|b| \geq ۵$ تعریف شده‌اند به طوری که با افزایش فراوانی تغییر در تنکی، $|a|$ یا $|b|$ از لحاظ اندازه افزایش می‌یابند. این نتایج به ناپایداری مدل برای داده‌های تولید شده با $|a| \geq ۵$ و $|b| \geq ۵$ اشاره می‌کند. در چنین مواقعی از معیارهای تأثیر معرفی شده انتظار می‌رود شروع به علامت‌گذاری مشاهده ۱ کنند.



شکل ۲. نسبت دفعات تغییر در تنکی برای $\hat{\beta}^{\text{lasso}}_{۱۰۰}$ (p) به ازای یک ترکیب معین از a و b به هنگام حذف مشاهده ۱

داده‌ها برای بررسی اثربخشی معیارهای تأثیر معرفی شده با استفاده از محدوده‌ای از مقادیر برای a و b تولید می‌شوند. برای هر ترکیب a و b، ۱۰۰۰ مجموعه داده تولید شدند. برای هر مجموعه داده ثبت گردید که آیا معیارهای تأثیر معرفی شده، مشاهدات بالقوه موثر را علامت‌گذاری کردند یا خیر. همچنین، تعداد مشاهدات نرمال یا غیرموثر

علامت‌گذاری شده ثبت شدند. از مقادیر بُرینش ۲± حاصل شده از لحاظ نظری استفاده می‌شود برای تعیین اینکه آیا مشاهده‌ای علامت‌گذاری می‌شود یا خیر. نتایج شبیه‌سازی در جدول ۱ برای محیط نامتعامد آورده می‌شود.

جدول ۱. نسبت دفعات تشخیص یک مشاهده موثر توسط معیارهای تأثیر لاسو در محیط طراحی نامتعامد

df-cvpath		df-regpath		df-lambda		df-model				
۲-۵۰	۱	۲-۵۰	۱	۲-۵۰	۱	۲-۵۰	۱	کل	b	a
۰/۰۰	۱/۰۰	۰/۰۳	۱/۰۰	۰/۰۵	۰/۳۵	۰/۰۳	۰/۸۱	۱/۰۰	۱۰	۱۰
۰/۰۶	۰/۰۰	۰/۰۵	۰/۰۰	۰/۰۵	۰/۰۱	۰/۰۵	۰/۰۱	۰/۰۲	۰	۱۰
۰/۰۰	۱/۰۰	۰/۰۲	۱/۰۰	۰/۰۴	۰/۶۷	۰/۰۳	۰/۸۶	۱/۰۰	۱۰	۰
۰/۰۳	۰/۹۳	۰/۰۳	۰/۹۶	۰/۰۵	۰/۳۳	۰/۰۴	۰/۶۰	۱/۰۰	۵	۵
۰/۰۶	۰/۰۰	۰/۰۵	۰/۰۰	۰/۰۵	۰/۰۲	۰/۰۵	۰/۰۱	۰/۰۲	۰	۵
۰/۰۳	۰/۹۱	۰/۰۳	۰/۹۸	۰/۰۵	۰/۳۷	۰/۰۴	۰/۶۱	۱/۰۰	۵	۰
۰/۰۵	۰/۰۷	۰/۰۵	۰/۰۴	۰/۰۵	۰/۱۰	۰/۰۵	۰/۰۸	۰/۱۹	۲	۲
۰/۰۶	۰/۰۰	۰/۰۵	۰/۰۰	۰/۰۵	۰/۰۱	۰/۰۵	۰/۰۱	۰/۰۱	۰	۲
۰/۰۵	۰/۰۶	۰/۰۵	۰/۰۴	۰/۰۵	۰/۱۰	۰/۰۵	۰/۰۸	۰/۲۰	۲	۰

شبیه‌سازی‌ها نشان می‌دهند که معیارهای معرفی شده در علامت‌گذاری مشاهده ۱ در مواقعی که موثر است، تأثیرگذار هستند. لازم به ذکر است در جدول ۱ "کل" نسبت دفعاتی را نشان می‌دهد که حداقل یکی از چهار معیار معرفی شده مشاهده ۱ را علامت‌گذاری کرد. زمانی که مشاهده ۱ در ناحیه تعریف شده با $|a| \geq ۵$ و $|b| \geq ۵$ قرار داشت، حداقل یکی از معیارها با فراوانی بالا مشاهده ۱ را علامت‌گذاری کرد. این ناحیه تأثیر[1] بود که در شکل ۲ نشان داده شد. به همین ترتیب، معیارهای معرفی شده این مشاهده را در مواقعی که $|a| = ۲$ و $|b| = ۲$ با درصد دفعات بسیار کوچک‌تری علامت‌گذاری کردند. این قابل انتظار است زیرا داده‌ها با واریانس ۱ تولید می‌شوند پس مقادیر $|a| = ۲$ و $|b| = ۲$ مربوط به مشاهده ۱ به طور خفیف تأثیرگذار است. Df-model مشاهده مندرج را زمانی که $a = ۱۰$ و $b = ۱۰$، ٪۸۱ از مواقع، $a = ۵$ و $b = ۵$، ٪۶۰ درصد از مواقع و $a = ۲$ و $b = ۲$، ٪۸ از مواقع علامت‌گذاری می‌کند. به طور کلی، این مقادیر با درصدهای تغییر در تنکی مشاهده شده در شکل ۲ مطابقت دارند و شواهدی ارائه می‌دهند که df-model مشاهده ۱ را به طور مناسب علامت‌گذاری می‌کند.

همچنین، نتایج شکل ۲ نشان می‌دهد که df-cvpath و df-regpath مشاهده ۱ را با درصد دفعات مناسبی علامت‌گذاری می‌کنند. معیار df-regpath تمایل دارد تا مشاهده ۱ را با درصد دفعات بزرگ‌تری نسبت به df-model علامت‌گذاری کند. این قابل انتظار است چون که df-regpath معیار حساس‌تری در مقایسه با df-model است. به خصوص، df-regpath تغییر واقعی در برآورد هر ضریب را اندازه‌گیری می‌کند در حالی که df-model فقط تغییرات تنکی در برآورد هر ضریب را تشخیص می‌دهد. Df-cvpath مشاهده مندرج را هنگامی که $|b| \geq ۵$، نزدیک به ٪۱۰۰ مواقع علامت‌گذاری می‌کند؛ شرایطی که انتظار می‌رود مشاهده مندرج تأثیر بزرگی روی خطای اعتبارسنجی متقابل داشته باشد. در نهایت، df-lambda مشاهده مندرج را با درصد دفعات کمتری نسبت به سایر معیارها علامت‌گذاری می‌کند. رفتار df-lambda احتمالاً بازتابی از تنوع ذاتی در انتخاب پارامتر منظم‌سازی برای

[1] Region of influence

رگرسیون لاسو است.

مهم‌تر از همه، شبیه‌سازی‌ها نشان می‌دهند که معیارهای تأثیر معرفی شده برای مشاهدات غیرموثر به درستی کار می‌کنند به این ترتیب که از این مشاهدات به طور متوسط تقریباً ۵٪ درصد یا کمتر علامت‌گذاری می‌شوند. این مطلوب است با توجه به اینکه مقادیر بُرینش نظری استفاده شده بر اساس خطای نوع اوّل ۵٪ هستند. علاوه بر این، هنگامی که مشاهده مندرج قدرت نفوذ بالایی دارد (مقدار بزرگ $|a|$) اما موثر نیست ($b = ۰$)، معیارهای معرفی شده به ندرت مشاهده ۱ را علامت‌گذاری می‌کنند.

نسبت دفعات تشخیص مشاهده موثر توسط معیارهای تأثیر تحت روش الاستیک‌نت به ازای αهای برابر ۰٫۷۵، ۰٫۵۰ و ۰٫۲۵ جهت انجام مقایسه در جدول‌های ۲، ۳ و ۴ برای محیط نامتعامد ارائه می‌شوند. نتایج شبیه‌سازی‌ها نشان می‌دهند که معیارهای تأثیر تحت الاستیک‌نت تأثیرگذار هستند در علامت‌گذاری مشاهده ۱ مواقعی که موثر است. لازم به ذکر است که معیارهای تشخیصی تأثیر برای الاستیک‌نت با $\alpha = ۰٫۷۵$ نسبت به الاستیک‌نت با $\alpha = ۰٫۵۰$ و $\alpha = ۰٫۲۵$ بهتر عمل می‌کنند.

جدول ۲. نسبت دفعات تشخیص یک مشاهده موثر توسط معیارهای تأثیر تحت روش الاستیک‌نت با $\alpha = ۰٫۷۵$ در محیط طراحی نامتعامد

df-cvpath		df-regpath		df-lambda		df-model				
۲-۵۰	۱	۲-۵۰	۱	۲-۵۰	۱	۲-۵۰	۱	کل	b	a
۰٫۰۰	۱٫۰۰	۰٫۰۳	۱٫۰۰	۰٫۰۵	۰٫۳۲	۰٫۰۳	۰٫۸۴	۱٫۰۰	۱۰	۱۰
۰٫۰۶	۰٫۰۰	۰٫۰۵	۰٫۰۰	۰٫۰۵	۰٫۰۱	۰٫۰۵	۰٫۰۱	۰٫۰۱	۰	۱۰
۰٫۰۰	۱٫۰۰	۰٫۰۲	۱٫۰۰	۰٫۰۴	۰٫۶۶	۰٫۰۳	۰٫۸۸	۱٫۰۰	۱۰	۰
۰٫۰۳	۰٫۹۰	۰٫۰۴	۰٫۹۴	۰٫۰۵	۰٫۳۳	۰٫۰۴	۰٫۶۱	۰٫۹۹	۵	۵
۰٫۰۶	۰٫۰۰	۰٫۰۵	۰٫۰۰	۰٫۰۵	۰٫۰۱	۰٫۰۵	۰٫۰۱	۰٫۰۱	۰	۵
۰٫۰۳	۰٫۹۰	۰٫۰۴	۰٫۹۴	۰٫۰۵	۰٫۳۳	۰٫۰۴	۰٫۶۱	۰٫۹۹	۵	۰
۰٫۰۵	۰٫۰۵	۰٫۰۵	۰٫۰۲	۰٫۰۵	۰٫۱۰	۰٫۰۵	۰٫۰۷	۰٫۱۶	۲	۲
۰٫۰۶	۰٫۰۰	۰٫۰۵	۰٫۰۰	۰٫۰۵	۰٫۰۱	۰٫۰۵	۰٫۰۱	۰٫۰۱	۰	۲
۰٫۰۶	۰٫۰۴	۰٫۰۵	۰٫۰۲	۰٫۰۵	۰٫۱۰	۰٫۰۵	۰٫۰۶	۰٫۱۶	۲	۰

جدول ۳. نسبت دفعات تشخیص یک مشاهده موثر توسط معیارهای تأثیر تحت روش الاستیک‌نت با $\alpha = ۰٫۵۰$ در محیط طراحی نامتعامد

df-cvpath		df-regpath		df-lambda		df-model				
۲-۵۰	۱	۲-۵۰	۱	۲-۵۰	۱	۲-۵۰	۱	کل	b	a
۰٫۰۱	۱٫۰۰	۰٫۰۳	۱٫۰۰	۰٫۰۵	۰٫۳۰	۰٫۰۳	۰٫۸۷	۱٫۰۰	۱۰	۱۰
۰٫۰۶	۰٫۰۰	۰٫۰۵	۰٫۰۰	۰٫۰۵	۰٫۰۱	۰٫۰۵	۰٫۰۰	۰٫۰۱	۰	۱۰
۰٫۰۱	۱٫۰۰	۰٫۰۳	۱٫۰۰	۰٫۰۴	۰٫۶۱	۰٫۰۳	۰٫۹۱	۱٫۰۰	۱۰	۰
۰٫۰۴	۰٫۸۲	۰٫۰۴	۰٫۷۸	۰٫۰۵	۰٫۲۷	۰٫۰۴	۰٫۵۹	۰٫۹۶	۵	۵
۰٫۰۶	۰٫۰۰	۰٫۰۵	۰٫۰۰	۰٫۰۵	۰٫۰۱	۰٫۰۵	۰٫۰۰	۰٫۰۱	۰	۵
۰٫۰۴	۰٫۸۳	۰٫۰۴	۰٫۸۰	۰٫۰۵	۰٫۳۱	۰٫۰۴	۰٫۶۰	۰٫۹۷	۵	۰
۰٫۰۶	۰٫۰۲	۰٫۰۵	۰٫۰۰	۰٫۰۵	۰٫۰۹	۰٫۰۵	۰٫۰۵	۰٫۱۲	۲	۲
۰٫۰۶	۰٫۰۰	۰٫۰۵	۰٫۰۰	۰٫۰۵	۰٫۰۱	۰٫۰۵	۰٫۰۰	۰٫۰۱	۰	۲
۰٫۰۶	۰٫۰۲	۰٫۰۵	۰٫۰۰	۰٫۰۵	۰٫۰۸	۰٫۰۵	۰٫۰۴	۰٫۱۱	۲	۰

جدول ۴. نسبت دفعات تشخیص یک مشاهده موثر توسط معیارهای تأثیر تحت روش الاستیک‌نت با $\alpha = ۰/۲۵$ در محیط طراحی نامتعامد

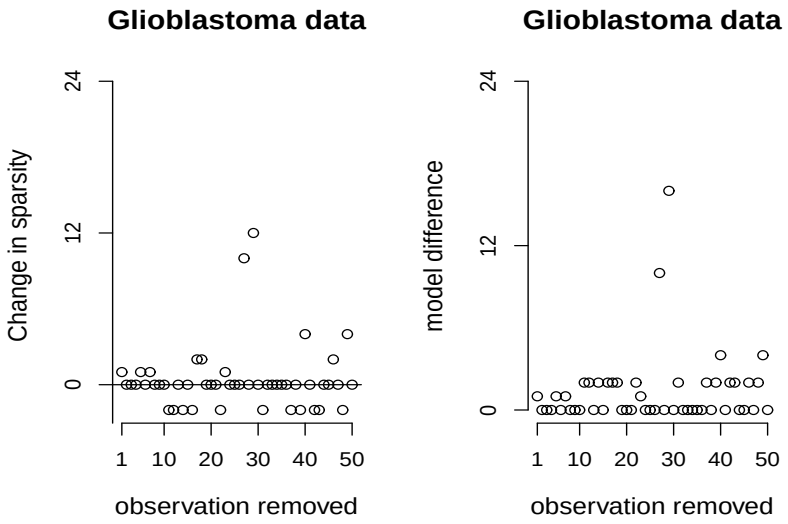
df-cvpath		df-regpath		df-lambda		df-model				
۲-۵۰	۱	۲-۵۰	۱	۲-۵۰	۱	۲-۵۰	۱	کل	b	a
۰/۰۲	۱/۰۰	۰/۰۵	۰/۹۴	۰/۰۵	۰/۲۹	۰/۰۴	۰/۹۵	۱/۰۰	۱۰	۱۰
۰/۰۷	۰/۰۰	۰/۰۶	۰/۰۰	۰/۰۵	۰/۰۱	۰/۰۵	۰/۰۰	۰/۰۱	۰	۱۰
۰/۰۲	۱/۰۰	۰/۰۵	۰/۹۸	۰/۰۵	۰/۳۹	۰/۰۴	۰/۹۶	۱/۰۰	۱۰	۰
۰/۰۵	۰/۳۴	۰/۰۶	۰/۰۱	۰/۰۵	۰/۱۲	۰/۰۴	۰/۳۰	۰/۵۳	۵	۵
۰/۰۷	۰/۰۰	۰/۰۶	۰/۰۰	۰/۰۵	۰/۰۱	۰/۰۵	۰/۰۰	۰/۰۱	۰	۵
۰/۰۵	۰/۳۴	۰/۰۵	۰/۰۱	۰/۰۵	۰/۱۲	۰/۰۴	۰/۳۰	۰/۵۳	۵	۰
۰/۰۶	۰/۰۰	۰/۰۶	۰/۰۰	۰/۰۵	۰/۰۳	۰/۰۵	۰/۰۰	۰/۰۳	۲	۲
۰/۰۷	۰/۰۰	۰/۰۶	۰/۰۰	۰/۰۵	۰/۰۱	۰/۰۵	۰/۰۰	۰/۰۱	۰	۲
۰/۰۶	۰/۰۰	۰/۰۶	۰/۰۰	۰/۰۵	۰/۰۴	۰/۰۵	۰/۰۰	۰/۰۴	۲	۰

۵ داده‌های توصیف ژن گلیوبلاستما

مجموعه داده از مطالعه توصیف ژن ریزآرایه گلیوبلاستما[1] حاصل شده است. میانگین زمان زنده ماندن بیماران مبتلا به گلیوبلاستما، شایع‌ترین تومور بدخیم مغزی اوّلیه در میان بزرگسالان، ۱۵ ماه است. هُروث و همکاران (۲۰۰۶) داده‌های توصیف ژن جهانی را برای ۳۶۰۰ ژن از ۱۲۰ بیمار گلیوبلاستما به دست آوردند (گروه ۱ = ۵۵ بیمار و گروه ۲ = ۶۵ بیمار). در این مقاله، تجزیه و تحلیل به داده‌های مربوط به گروه ۱ محدود می‌شود. مشابه تحلیل این داده‌ها توسط وانگ و همکاران (۲۰۱۱) پنج بیمار از گروه ۱ که در آخرین پیگیری هنوز زنده بودند، حذف شدند. همچنین، لگاریتم زمان زنده ماندن به عنوان متغیر پاسخ استفاده شد. مجموعه کامل ۳۶۰۰ ژن در تجزیه و تحلیل گنجانده شد به این معنی که این مثال با وضعیت $n = ۵۰$ و $p = ۳۶۰۰$ مطابقت دارد (راجاراتنام و همکاران ، ۲۰۱۹). تغییر در تنکی[2] و تفاوت مدل برای بررسی تأثیر روی راه‌حل لاسو به هنگام حذف مشاهدات فردی مشاهده می‌شوند. تعداد برآوردهای ضریب صفر، تنکی نامیده می‌شود. تفاوت مدل، تعداد تفاوت‌ها در شمول و عدم شمول متغیرهای کمکی بین راه‌حل لاسو بر اساس مجموعه داده کامل و راه‌حل لاسو با مشاهده خاص حذف شده، تعریف می‌شود. به عنوان مثال، اگر حذف یک مشاهده منجر به تغییر متغیرهای کمکی موجود در مدل انتخاب شده از $X_۳$، $X_۴$ و $X_۵$ به $X_۱$، $X_۲$ و $X_۳$ گردد، تغییر در تنکی و تفاوت مدل به ترتیب برابر ۰ و ۴ خواهند بود.

شکل ۳، تغییر در تنکی و تفاوت مدل را برای مجموعه داده گلیوبلاستما نشان می‌دهد. این شکل به وضوح نشان می‌دهد که عدم وجود مشاهدات فردی می‌تواند به معنای تغییر معنی‌داری در راه‌حل لاسو باشد. به عنوان مثال، اگر مشاهده ۲۷ یا ۲۹ در داده‌های گلیوبلاستما وجود نداشت، تغییر در تنکی به ترتیب برابر ۱۰ و ۱۲ مشاهده می‌شد. بنابراین، تأثیرات چنین تغییر بزرگی در مدل انتخاب شده توسط لاسو به طور بالقوه بسیار معنی‌دار است. به ویژه، اگر نتایج به معنای واقعی کلمه در نظر گرفته شوند، می‌تواند منجر به بررسی بعدی این ۱۰ یا ۱۲ ژن بر اساس داده‌های

[1] Glioblastoma microarray gene expression
[2] Sparsity



شکل ۳. تغییر در تُنُکی و تفاوت مدل راه‌حل لاسو برای داده‌های گلیوبلاستما

تازه به دست آمده شود (در مقایسه با یافته‌ای که تحقیقات بیشتر بی‌دلیل است). مطابق شکل ۳ به هنگام حذف مشاهده ۲۷ یا ۲۹ تفاوت مدل به ترتیب برابر ۱۰ و ۱۶ است. لازم به ذکر است که ویژگی‌های مجموعه داده مورد بررسی در اینجا برای برجسته کردن حساسیت بالقوه لاسو به مشاهدات فردی انتخاب شدند. این مثال به وضوح تأکید می‌کند بر نیاز به معیارهایی برای سنجش مشاهداتی که پتانسیل اعمال تأثیر قوی بر راه‌حل لاسو را دارند.

با بیش از ۱۰۰ تخصیص تصادفی مشاهدات به دسته‌ها، تعداد متغیرهای کمکی برای لاسو (الاستیک‌نت با $\alpha = ۱$) و الاستیک‌نت با αهای برابر ۰٫۷۵، ۰٫۵۰ و ۰٫۲۵ به ترتیب از ۰ تا ۵۵، ۰ تا ۵۰، ۰ تا ۷۱ و ۰ تا ۱۰۳ ژن متغیر بودند. می‌توان نتیجه گرفت مدل انتخاب شده پس از حذف مشاهدات موثر و برازش الاستیک‌نت تحت تأثیر تخصیص مشاهدات به دسته‌های داده‌ها در اعتبارسنجی متقابل قرار دارد. لذا، اعمال الاستیک‌نت بر روی چندین تخصیص متفاوت مشاهدات به دسته‌ها (m) به عنوان تحلیلی برای داده‌های گلیوبلاستما است. نسبت دفعاتی که هر ژن در m مدل انتخاب شده قرار می‌گیرد، احتمال شمول نامیده می‌شود که اهمیت نسبی هر ژن را نشان می‌دهد. ژن‌هایی با احتمال شمول ۱۰-برتر تحت الاستیک‌نت با αهای برابر ۱، ۰٫۷۵، ۰٫۵۰ و ۰٫۲۵ پس از حذف مشاهدات موثر روی $m = ۱۰۰$ تخصیص متفاوت مشاهدات به دسته‌ها در جدول ۵ ارائه می‌شود. در خروجی جدول ۵ ژن‌هایی با احتمال شمول بالا برای الاستیک‌نت با $\alpha = ۱$ (لاسو) نظیر ژن‌هایی با احتمال شمول بالا برای الاستیک‌نت با $\alpha = ۰٫۷۵$ هستند. برترین ژن KCNC1 است که در ۸۱ مدل از ۱۰۰ مدل انتخاب شده توسط لاسو قرار داشت. همچنین، این ژن در ۹۰ مدل از ۱۰۰ مدل انتخاب شده توسط الاستیک‌نت با $\alpha = ۰٫۷۵$ قرار داشت.

جدول ۵. ژن‌هایی با احتمال شمول ۱۰–برتر بعد از حذف مشاهدات موثر

احتمال شمول تحت الاستیک‌نت				
$\alpha = ۰/۲۵$	$\alpha = ۰/۵۰$	$\alpha = ۰/۷۵$	$\alpha = ۱$	ژن
۰/۹۵	۰/۹۶	۰/۹۰	۰/۸۱	KCNC1
۰/۹۲	۰/۹۲	۰/۸۶	۰/۷۸	PTEN
۰/۸۹	۰/۹۲	۰/۸۵	۰/۷۷	SYNJ2
۰/۹۴	۰/۹۲	۰/۸۵	۰/۷۶	CNN3
۰/۹۲	۰/۸۶	۰/۷۹	۰/۷۰	FLJ12443
۰/۹۲	۰/۸۸	۰/۸۲	۰/۷۰	CGI-115
۰/۹۱	۰/۸۷	۰/۷۳	۰/۶۲	PTGDS
۰/۷۶	۰/۶۹	۰/۶۴	۰/۶۰	IRF3
۰/۷۶	۰/۸۲	۰/۶۴	۰/۶۰	GTSE1
۰/۶۹	۰/۷۵	۰/۵۹	۰/۶۰	ADIPOR1
۰/۳۹	۰/۳۳	۰/۶۰	۰/۵۱	EDN1
۰/۹۱	۰/۷۵	۰/۳۳	۰/۱۶	SLC31A2
۰/۸۳	۰/۶۳	۰/۰۹	۰/۰۹	CEBPD

احتمالات شمول برای همه ژن‌های ۱۰–برتر تحت لاسو بیشتر از ٪۵۰ هستند که می‌تواند نشان دهنده ارتباط هر یک از ژن‌ها با زنده ماندن بیمار باشد. متقابلاً، تحت الاستیک‌نت با $\alpha = ۰/۷۵$ همه ژن‌های ۱۰–برتر، احتمالات شمول بیشتر از ٪۵۰ دارند. تفاوت‌های جزئی بین احتمالات شمول لاسو و الاستیک‌نت با $\alpha = ۰/۷۵$ در جدول ۵ حاکی از شناسایی و حذف مشاهدات موثر یکسان (مشاهده ۲۷ و ۲۹) است. به ویژه، تحقیق در زمینه ادبیات مربوط به چهار مورد از ژن‌ها با بالاترین احتمالات شمول لاسو نشان می‌دهد که هر ژن ممکن است ارتباط بیولوژیکی مهمی داشته باشد (داساری و همکاران ، ۲۰۱۰؛ لیو و همکاران ، ۲۰۰۶؛ وبستر و همکاران ، ۲۰۰۹). برای مثال، ژن KCNC1 نشان داده شده است با زنده ماندن در گلیوبلاستما ارتباط دارد (لیو و همکاران ، ۲۰۰۶)، ژن PTEN یک سرکوب کننده معروف گلیوبلاستما و طیفی از سرطان‌های دیگر است (داساری و همکاران ، ۲۰۱۰) و ژن CNN3 اخیراً به عنوان یک ژن با اولویت بالقوه در تحقیقات سرطان شناخته شده است (وبستر و همکاران ، ۲۰۰۹). در جدول ۵ ژن‌های SLC31A2 و CEBPD احتمالات شمول بزرگی تحت الاستیک‌نت با $\alpha = ۰/۵۰$ و $\alpha = ۰/۲۵$ نسبت به لاسو و الاستیک‌نت با $\alpha = ۰/۷۵$ دارند که بیانگر عدم شناسایی و حذف مشاهده موثر ۲۷ قبل از برازش الاستیک‌نت با αهای مربوطه است. در طول محاسبه احتمالات شمول نیز می‌توان مشاهده کرد نسبت دفعاتی که هر مشاهده به عنوان مشاهده موثر علامت‌گذاری می‌شود. نسبت دفعات با مقادیر بالا برای یک مشاهده نشان دهنده تأثیر بزرگ آن مشاهده بر مدل برازش شده است. برای داده‌های گلیوبلاستما توسط لاسو و الاستیک‌نت با αهای برابر ۰/۷۵، ۰/۵۰ و ۰/۲۵ به ترتیب ۲، ۲، ۱ و ۱ مورد از ۵۰ مشاهده در بیش از ٪۵۰ مواقع به عنوان مشاهده موثر علامت‌گذاری شدند. لاسو مشاهدات ۲۷ و ۲۹ را به ترتیب در ٪۹۲ و ٪۱۰۰ مواقع علامت‌گذاری کرد. همچنین، الاستیک‌نت با $\alpha = ۰/۷۵$ مشاهدات ۲۷ و ۲۹ را به ترتیب در ٪۹۱ و ٪۱۰۰ مواقع علامت‌گذاری کرد اما فقط مشاهده ۲۹ برای الاستیک‌نت با $\alpha = ۰/۵۰$ و $\alpha = ۰/۲۵$ در ٪۱۰۰ مواقع علامت‌گذاری شد. لذا، الاستیک‌نت با $\alpha = ۰/۷۵$ مشابه لاسو مشاهدات ۲۷ و ۲۹ را در بیش از ٪۵۰ مواقع علامت‌گذاری کرد. این موضوع نشان می‌دهد

که عاقلانه است این مشاهدات برای ارزیابی دلیل تأثیر آنها بیشتر بررسی شوند. بررسی سطحی از داده‌ها نشان می‌دهد که مشاهده ۲۹ با ۷ روز کمترین زمان زنده ماندن را دارد و کوچک‌ترین مقدار بعدی ۴۳ روز است. علاوه بر این، مشاهده ۲۷ مجموعه‌ای نسبتاً دور از مقادیر توصیف (یا متغیر کمکی) را دارد. مقادیر متغیر کمکی مقیاس‌پذیر و متمرکز (در مورد انحراف استاندارد و میانگین) این مشاهده مشخص شدند که چهارمین فاصله اقلیدسی بزرگ از میانگین را دارند.

بحث و نتیجه‌گیری

دسترسی گسترده به داده‌های بعد بالا، استفاده از روش‌های درستنمایی تاوانیده را رایج ساخته است. در این مقاله، توانمندی معیارهای تأثیر معرفی شده تحت لاسو و الاستیک‌نت با αهای برابر ۰٫۷۵، ۰٫۵۰ و ۰٫۲۵ برای تشخیص مشاهدات موثر در داده‌های بعد بالا به طور نظری بررسی شده و به صورت عددی ارزیابی شده‌اند. همان‌طور که از نتایج شبیه‌سازی ملاحظه می‌شود، می‌توان گفت معیارهای تأثیر تحت الاستیک‌نت نظیر معیارهای تأثیر تحت لاسو در شناسایی مشاهدات موثر تأثیرگذار هستند. این ویژگی برای مشاهدات غیرموثر نیز برقرار است. همچنین، معیارهای تأثیر تحت الاستیک‌نت از طریق داده‌های واقعی عملکرد خوبی از خود نشان می‌دهند زیرا نتایج تعیین مشاهدات موثر و ژن‌های مهم پس از حذف مشاهدات موثر شناسایی شده توسط این معیارها و برازش مدل الاستیک‌نت به ویژه با $\alpha = ۰٫۷۵$ تفاوت چندانی با خروجی تحلیل تحت لاسو ندارد. بی‌شک نیاز برای معیارهای تشخیص مشاهدات موثر در محیط بعد بالا مهم‌تر از تحلیل OLS استاندارد است زیرا: (آ) هر مشاهده‌ای در تحلیل داده‌های بعد بالا می‌تواند علاوه بر برآورد پارامترها روی انتخاب مدل نیز در مقایسه با تحلیل OLS تأثیر بگذارد و (ب) هر تحلیلی در محیط بعد بالا که $p \gg n$ ذاتاً ناپایدارتر است به این معنی که پتانسیل تأثیرگذاری یک مشاهده از قبل در مقایسه با تحلیل OLS که $n > p$ به طور چشمگیری افزایش داده می‌شود.

تقدیر و تشکر

نویسندگان مقاله از پیشنهادات ارزنده داوران، سردبیر و ویراستار محترم در راستای بهبود سطح کیفی مقاله کمال تشکر و قدردانی را دارند.

مراجع

معنوی، م. و روزبه، م. (۱۳۹۹)، روش‌های تحلیل رگرسیونی برای داده‌های بعد بالا، مجله اندیشه آماری، **۲۵**(۱)، ۶۹-۹۰.

نوری، ن. (۱۴۰۱)، تشخیص و آنالیز داده‌های موثر برای رگرسیون ابعاد بالا، پایان‌نامه کارشناسی ارشد، دانشگاه تبریز، تبریز.

Atkinson, A. C. (1981). Two Graphical Displays for Outlying and Influential Observations in Regression, *Biometrika*, **68**(1), 13-20.

Atkinson, A. C. (1984). Two Books on Regression Diagnostics. *Annals of Statistics*, **12**, 392-401.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

Chen, X., Wang, Z. J., and McKeown, M. J. (2010). Asymptotic Analysis of Robust LASSOs in the Presence of Noise with Large Variance. *IEEE Transactions on Information Theory*, **56**(10), 5131-5149.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, **19(1)**, 15-18.

Cook, R. D. (1979). Influential Observations in Linear Regression. *Journal of the American Statistical Association*, **74**(365), 169-174.

Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Dasari, V. R., Kaur, K., Velpula, K. K., Gujrati, M., Fassett, D., Klopfenstein, J. D., ... and Rao, J. S. (2010). Upregulation of PTEN in Glioma Cells by Cord Blood Mesenchymal Stem Cells Inhibits Migration via Downregulation of the PI3K/Akt Pathway. *PloS one*, **5**(4), e10350.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55-67.

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., ... and Mischel, P. S. (2006). Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Molecular Target. *Proceedings of the National Academy of Sciences*, **103**(46), 17402-17407.

Lambert-Lacroix, S., and Zwald, L. (2011). Robust Regression Through the Huber's Criterion and Adaptive Lasso Penalty. *Electronic Journal of Statistics*, **5**, 1015-1053.

Liu, F., Park, P. J., Lai, W., Maher, E., Chakravarti, A., Durso, L., ... and Johnson, M. D. (2006). A Genome-Wide Screen Reveals Functional Gene Clusters in the Cancer Genome and Identifies EphA2 as a Mitogen in Glioblastoma. *Cancer Research*, **66**(22), 10815-10823.

Rajaratnam, B., Roberts, S., Sparks, D., and Yu, H. (2019). Influence Diagnostics for High-Dimensional Lasso Regression. *Journal of Computational and Graphical Statistics*, **28**(4), 877-890.

She, Y., and Owen, A. B. (2011). Outlier Detection Using Nonconvex Penalized Regression. *Journal of the American Statistical Association*, **106**(494), 626-639.

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

Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random Lasso. *The Annals of Applied Statistics*, **5**(1), 468-485.

Wang, T., and Li, Z. (2017). Outlier Detection in High-Dimensional Regression Model. *Communications in Statistics-Theory and Methods*, **46**(14), 6947-6958.

Webster, R. J., Giles, K. M., Price, K. J., Zhang, P. M., Mattick, J. S., and Leedman, P. J. (2009). Regulation of Epidermal Growth Factor Receptor Signaling in Human Cancer Cells by MicroRNA-7. *Journal of Biological Chemistry*, **284**(9), 5731-5741.

Zhao, J., Leng, C., Li, L., and Wang, H. (2013). High-Dimensional Influence Measure, *Annals of Statistics*, **41**, 2639-2667.

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