بسم الله الرحمن الرحیم

عنوان: تشخیص مشاهدات مؤثر برای رگرسیون با ابعاد بالا

نام: احمدرضا ملک محمودی

استاد راهنما: دکتر صابری

فهرست مطالب

فصل اول:

مفاهیم مقدماتی

در این فصل، مفاهیم ابتدایی مرتبط با داده‌های با ابعاد بالا و کاربرد روش‌های تاوانیده در تحلیل‌های رگرسیونی مورد بررسی قرار می‌گیرند.

۱ روش‌های تاوانیده در رگرسیون

روش‌های تاوانیده در رگرسیون زمانی به کار می‌روند که تعداد متغیرهای توضیحی (p) بسیار بیشتر از تعداد مشاهدات (n) باشد، یا وقتی بخواهیم تأثیر برخی متغیرهای اضافی را کاهش دهیم. این روش‌ها با افزودن یک جمله تاوان به تابع هدف، مانع از پیچیدگی بیش از حد مدل و کاهش هم‌خطی میان متغیرها می‌شوند. از جمله رایج‌ترین این روش‌ها می‌توان به رگرسیون لاسو (Lasso) و ریج (Ridge) اشاره کرد.

اولین و ساده ترین روش انتخاب متغیر، انتخاب بهترین زیر مجموعه ممکن است که در آن با در اختیار داشتن k متغیر رگرسیونی، تمام β مدل زیر مجموعه را برازش می دهیم (با فراض وجود β در همه مدل ها) سپس بهترین مدل را از بین آنها انتخاب می کنیم.

این روش دارای چند ایراد است :

1_بار محاسباتی آن زیاد است.

2_فرایند انتخاب متغیر و برآورد پارامتر ها به صورت جداگانه انجام می شود و این امر باعث می شود که مدل حاصل از آن، ناپایدار باشد به طوری که با تغییر کوچکی در مشاهدات نمونه، مدل انتخاب شده تغییر کند.

به دلیل ایراد ایراد اول، الگوریتم هایی مانند پسرو و... معرفی شدند که آنها هم ایراداتی دارند؛ ازجمله همان ایراد دوم بالا ،اینکه خطای حاصل از هرقدم به قدم های بعدی منتقل می شوند.

از طرفی، زمانی که $n < p$ باشد ،روش حذف پسرو قادر به برازش مدل اولیه با همه ی متغیر های رگرسیونی موجود نیست چون در این حالت$(\underline{X'X})$معکوس پذیر نخواهد بود.

روش پیشرو و قدم به قدم هم نهایتا اجازه ورود n متغیر را به مدل می دهند و بعد از آن مسئاله بالا رخ می دهد.

به عنوان یک راه حل برای ایراد های بالا، روشهای انقباضی (shrinkage) معرفی شده اند. از روش های براورد پارامتر انقباضی می توان به رگرسیون های ریج و لاسو اشاره کرد.

Lasso: Least absolute Shrinkage and Selection Operator

ویژگی‌ها و کاربردها:

1. **حذف متغیرهای غیرمرتبط:** برخی ضرایب را به‌طور دقیق برابر صفر می‌کند.
2. **کاهش پیچیدگی مدل:** مانع از بیش‌برازش (Overfitting) می‌شود.
2. **انتخاب متغیرها:** متغیرهایی که بیشترین تأثیر را دارند شناسایی می‌شوند.

۲ رگرسیون لاسو

رگرسیون لاسو توسط تیبشیرانی (1996) به عنوان تعمیمی از روش حداقل مربعات معمولی (OLS) معرفی شد. این روش با افزودن یک جریمه ℓ_1 به تابع هدف، برخی ضرایب را دقیقاً برابر صفر می‌کند و در نتیجه یک مدل تنک تولید می‌کند.

تعریف

تابع هدف در لاسو به صورت زیر تعریف می‌شود:

$$\hat{\underset{\sim}{\beta}} = \arg\min_{\underset{\sim}{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\} + \lambda \sum_{j=1}^{p} |\beta|_j \qquad (۱)$$

λ : پارامتر تنظیم یا تاوان که تعادل میان برازش و پیچیدگی مدل را تعیین می‌کند.

β : ضرایب رگرسیون که باید برآورد شوند.

λ : مقدار بزرگتر از صفر، برخی ضرایب را صفر کرده و متغیرهای غیرمؤثر را حذف می‌کند.

۲.۱ برآورد پارامترها و انتخاب متغیرها

در رگرسیون لاسو، به دلیل استفاده از جریمه ℓ_1 ، برخی از ضرایب به سمت صفر منقبض می‌شوند.

فرآیند انتخاب متغیر از طریق مقایسه مقادیر جریمه و اثر هر متغیر بر پاسخ انجام می‌شود. انتخاب λ معمولاً با استفاده از روش‌های اعتبارسنجی متقابل یا معیارهای اطلاعاتی انجام می‌شود.

۲.۲ معیارهای تأثیر در لاسو

معیارهای مختلفی برای ارزیابی تأثیر مشاهدات در لاسو معرفی شده‌اند که هرکدام جنبه خاصی از مدل‌سازی را هدف قرار می‌دهند. این معیارها عبارتند از:

1. **معیار df-Model :** تغییر در مدل انتخابی هنگام حذف یک مشاهده.
2. **معیارdf-Regpath :** تغییر در مسیر تنظیم لاسو هنگام حذف یک مشاهده.
3. **معیارdf-Cvpath :** تغییر در خطای پیش‌بینی در مسیر اعتبارسنجی متقابل.
4. **معیارdf-Lambda :** تغییر در مقدار بهینه λ هنگام حذف یک مشاهده.

تمامی این معیارها در ادامه مقاله با جزئیات کامل بررسی خواهند شد.

۳ معیارهای تأثیر برای رگرسیون لاسو

۳.۱ معیار Df-Model

اوّلین معیار تأثیری که معرفی می شود، df -model است که به طور مستقیم تغییر در مدل انتخاب شده توسط لاسو را زمانی که مشاهده ای حذف شده، ارزیابی می کند. تعیین اندازه این تغییر مهم است چون یک تغییر بزرگ در مدل انتخاب شده توسط لاسو می تواند به طور چشمگیری نتایج حاصل از تجزیه و تحلیل را تغییر دهد .Df - model برای i ـ امین مشاهده (راجاراتنام و همکاران ، ۲۰۱۹) به صورت

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

تعریف شده است که $\delta(i)=\sum_{j=1}^{p}\left|I\left\{\hat{\beta}_j^{\,lasso}=0\right\}-I\left\{\hat{\beta}_j^{lasso}(i)=0\right\}\right|$ و $I\{.\}$ تابع نشانگر را نشان می دهد.

به ویژه، df-model یک معیار مقیاس شده از تعداد تغییرات در متغیرهای پیشگوی انتخاب شده است که در راه حل لاسو به هنگام حذف یک مشاهده روی می دهد.

محاسبه df-model شامل برازش ۱ $+n$ بار لاسو است و پس از آن مقادیر مشاهده شده $\delta(i)$ حساب می شوند. میانگین نمونه و واریانس n مقدار مشاهده شده $\delta(i)$ می توانند به ترتیب به عنوان برآوردهای $E\{\delta(i)\}$ و $Var\{\delta(i)\}$ به کار روند. نظریه ای که در ادامه اثبات می کند، مقادیر محاسبه شده df - model را می توان با مقادیر بُرینش ۲± مقایسه کرد.

نتیجه مقادیر بُرینش برای df - model در محیط طرح متعامد بعد بالا نشان داده خواهد شد. این نتیجه مقادیر بُرینش طبیعی و قدرمطلق مشاهداتی با بزرگی (i)df – model بیش از ۲ را می دهد که می توان برای بررسی بیشتر نشانه گذاری کرد.

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

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

برای هر $i\in\{1,...,n\}$ 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 شده اند تا میانگین ۰ و واریانس ۱ داشته باشند و خطاهای ε_i متغیرهای تصادفی مستقل با توزیع $N(0,\ \sigma^2)$ که $\sigma^2>0$ مجهول است. حال فرض می شود

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

همچنین فرض می شود $\hat{\beta}_j^{lasso}(s)$ ، برآوردگر لاسو برای ضریب j ـ ام را نشان می دهد زمانی که پارامتر جریمه ℓ_1 در درستنمایی مقدار s می گیرد. اگر دنباله ای از مقادیر ضرایب واقعی صدق کند در

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

آنگاه $p \to \infty$ (با $n \equiv n_p \to \infty$). بنابراین دنباله ای از متغیرهای تصادفی $\Delta \equiv \Delta_p$ صدق می کند در

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

۳.۲ معیار Df-Regpath

حال معیار تأثیر df-regpath معرفی می شود تا تغییر کلی در مسیر منظم سازی راه حل لاسو هنگامی که یک مشاهده معین حذف شده، سنجیده شود. ارزیابی این تغییر مهم است زیرا یک تغییر بزرگ در مسیر منظم سازی به این معنی است که نمایه ضرایب برآورد شده لاسو به دلیل حذف یک مشاهده به طور قابل توجهی تغییر کرده است. چنین تغییراتی می تواند به این معنی باشد که مشاهده موردنظر تأثیر قابل ملاحظه ای روی ضرایب برآورد شده لاسو و به طور بالقوه تغییر در تفسیر و نتایج حاصل از تجزیه و تحلیل لاسو دارد. Df-regpath برای i- امین مشاهده به صورت (راجاراتنام و همکاران ، ۲۰۱۹)

$$df \ - \ regpath(i) \ = \frac{\Delta_1 \hat{\underset{\sim}{\beta}}^{lasso}(i) - E\left\{\Delta_1 \hat{\underset{\sim}{\beta}}^{lasso}(i)\right\}}{\sqrt{Var\left\{\Delta_1 \hat{\underset{\sim}{\beta}}^{lasso}(i)\right\}}} \qquad (۳)$$

تعریف شده است، که در آن $\Delta_1 \hat{\underset{\sim}{\beta}}^{lasso}(i) = \int_l^u \left\| \hat{\underset{\sim}{\beta}}^{lasso}(s) - \hat{\underset{\sim}{\beta}}^{lasso}(s,i) \right\|_1 ds$ و l و u به گونه ای مشخص شده اند که بازه $[l,u]$ محدوده ای از مقادیر قابل قبول λ را تعریف می کند.به طور خاص، df-regpath یک معیار مقیاس شده از تغییر کلی در مسیر منظم سازی 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 دربر می گیرد. از این $n+1$ مدل لاسو برازش شده، نمایه های $\hat{\underset{\sim}{\beta}}^{lasso}(s)$ و $\hat{\underset{\sim}{\beta}}^{lasso}(s,i)$ به دست می آیند.سپس انتگرال $\Delta_1 \hat{\underset{\sim}{\beta}}^{lasso}(i)$ با استفاده از تکنیک های عددی استاندارد که برای مقادیر $\left\| \hat{\underset{\sim}{\beta}}^{lasso}(s) - \hat{\underset{\sim}{\beta}}^{lasso}(s,i) \right\|_1$ روی دنباله ای از مقادیر λ اعمال شده، تقریب می یابد. میانگین نمونه و واریانس n مقدار مشاهده شده $\Delta_1 \hat{\underset{\sim}{\beta}}^{lasso}(i)$ را می توان به ترتیب به عنوان برآوردهای $E\left\{\Delta_1 \hat{\underset{\sim}{\beta}}^{lasso}(i)\right\}$ و $Var\left\{\Delta_1 \hat{\underset{\sim}{\beta}}^{lasso}(i)\right\}$ استفاده کرد. دوباره مشابه df-model، نظریه ای که در ادامه نشان می دهد می توان مقادیر محاسبه شده df-regpath را با مقادیر بُرینش ۲± 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 می شود که تغییرپذیری نمونه گیری معیار تأثیر df-regpath را در محیط طراحی متعامد اندازه می گیرد.

قضیه ۲. مدل رگرسیونی تنک توصیف شده در قضیه ۱ را در نظر بگیرید و فرض کنید وجود دارد $\eta < 1$ ، $\varepsilon > 0$ و $M > 0$ به طوری که

$$\limsup_{p\to\infty}\left[\frac{1}{p}\sum_{j=1}^{p} I\left\{\beta_j^{true} \neq 0\right\}\right] = \eta$$

و برای همه $p \geq 1$ ،

$$\frac{1}{p}\sum_{j=1}^{p}\left|\beta_j^{true}\right|^{4+\varepsilon} \leq M. \qquad (۴)$$

آنگاه:

الف_ امید ریاضی $\mu_j(\lambda) = E\left[\hat{\beta}_j^{lasso}(\lambda)\right]$ برای هر $j \in \{1,...,p\}$ و برای همه $\lambda > 0$ متناهی است؛

ب_ انتگرال تصادفی

$$\Omega_p = \int_0^{\infty}\left\|\hat{\underset{\sim}{\beta}}^{lasso}(\lambda) - \underset{\sim}{\mu}(\lambda)\right\|_1 d\lambda = \int_0^{\infty}\sum_{j=1}^{p}\left|\hat{\beta}_j^{\,lasso}(\lambda) - \mu_j(\lambda)\right| d\lambda$$

برای همه بردار های 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_ $0 < E\left(\Omega_p\right) < \infty$ و $0 < Var\left(\Omega_p\right) < \infty$ ؛

د_ وقتی $p \to \infty$ ، $\frac{\Omega_p - E\left(\Omega_p\right)}{\sqrt{Var\left(\Omega_p\right)}} \hat{I} \quad N(0,1)$.

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

۳/۳ معیار Df-Cvpath

اکنون معیار تأثیر df-cvpath معرفی می شود تا تغییر در عملکرد پیشگویی لاسو هنگامی که مشاهده ای 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 راه حل لاسو و یک مقدار پاسخ غیرمعمول دارد. منحنی خطای اعتبارسنجی متقابل $\gamma(s)$ برای مقادیر متفاوت پارامتر منظم سازی $\lambda = s$ ، خطای پیشگویی روی داده های آزمون را بعد از اینکه روش لاسو روی مولفه متفاوت داده ها آموزش دیده، اندازه می گیرد. Df-cvpath برای i ـ امین مشاهده به صورت (راجاراتنام و همکاران ، ۲۰۱۹)

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

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 i ـ امین مشاهده حذف شده، است. l و u به گونه ای تعریف شده اند که بازه $[l,u]$ محدوده ای از مقادیر قابل قبول λ را معین می کند. به طور مشخص، df-cvpath یک معیار مقیاس شده از تغییر کلی در مسیر اعتبارسنجی متقابل لاسو زمانی که مشاهده ای حذف شده، است.

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

۳.۴ معیار Df-Lambda

معرفی می شود تا تغییر در مقدار بهینه پارامتر منظم سازی لاسو λ lambda-df حال معیار تأثیر زمانی که مشاهده ای حذف شده، سنجیده شود. محاسبه این تغییر مهم است زیرا تغییر بزرگی در مقدار بهینه λ نشان می دهد که مشاهده موردنظر تأثیر زیادی بر میزان کاهش برآوردهای ضرایب مدل انتخابی لاسو دارد.Df-lambda برای i ـ امین مشاهده به صورت (راجاراتنام و همکاران ، ۲۰۱۹)

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

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 i ـ امین 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 $\hat{\lambda}-\hat{\lambda}(i)$ می توان به ترتیب به عنوان برآوردهای $E\left\{\hat{\lambda}-\hat{\lambda}(i)\right\}$ و $Var\left\{\hat{\lambda}-\hat{\lambda}(i)\right\}$ استفاده کرد.

مقادیر بُرینش ±۲ برای df-lambda را می توان از نظر اکتشافی توجیه کرد. برای k های بزرگ که k تعداد دسته ها در اعتبارسنجی متقابل است، خطای اعتبارسنجی متقابل به صورت نرمال توزیع شده است. این امر با تشخیص خطای اعتبارسنجی متقابل $\gamma\left(\hat{\lambda}\right)=\frac{1}{k}\sum_{i=1}^{k}PE_i\left(\hat{\lambda}\right)$ (که $PE_i\left(\hat{\lambda}\right)$ خطای پیشگویی روی iـ امین دسته از داده ها را نشان می دهد) که یک میانگین است، دنبال می شود و پس از آن می توان قضیه حد مرکزی را اعمال کرد. فراخوانی روش دلتا و معکوس کردن $\gamma\left(\hat{\lambda}\right)$ تا $\hat{\lambda}$ به دست آید. همچنین نتیجه گرفته می شود که $\hat{\lambda}$ تقریباً به صورت نرمال با میانگین و واریانس متناظر توزیع شده است. بنابراین، کمیت استاندارد شده در معادله (۶) را می توان با مقادیر بُرینش تقریبی ±۲ استفاده کرد.

۴ رگرسیون الاستیک‌نت

از محدودیت های لاسو می توان به موارد زیر اشاره کرد،

الف_ اگر $p>n$ باشد، لاسو حداکثر n متغیر می تواند 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 مجموعه ای از متغیرها رو داشته باشیم که دارای هم خطی بالایی باشند، لاسو فقط یک متغیر آن هم به صورت تصادفی انتخاب کرده و از بقیه متغیرهای مجموعه چشم پوشی می کند که این برای تکرارپذیری و تفسیر داده ها خوب نیست.

با توجه به محدودیت های لاسو، رگرسیون الاستیک‌نت که ترکیبی از رگرسیون ریج (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 بهینه سازی به صورت

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

تعریف می شود. 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 روش هایی که تنها یک پارامتر تاوان دارند، امری به

مراتب دشوارتر خواهد بود و این یکی از معایب این روش محسوب می شود. با در نظر گرفتن $\alpha = \lambda_1/(\lambda_1 + \lambda_2)$ می توان برآوردگر الاستیک‌نت را به فرم تاوانی

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

نوشت. ناحیه تاوان در این روش ترکیبی، اکیداً محدب از نواحی تاوان در روش های 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 می توان برای گروه بندی متغیرهای پیشگو استفاده کرد به خصوص زمانی که تعداد آنها زیاد است. برآوردگر اصلاح شده الاستیک‌نت که دارای دقت پیش بینی بالاتری نسبت به برآوردگر اوّلیه است به صورت $\hat{\beta}^{ENet} = (1+\lambda_2)\hat{\beta}$ تعریف می شود که در آن $\hat{\beta}^{ENet}$ و $\hat{\beta}$ به ترتیب برآوردگرهای الاستیک‌نت و الاستیک‌نت اوّلیه (خام) است.

فصل دوم:

شبیه‌سازی و نتایج

۵ مطالعه شبیه سازی

در این بخش، توانایی روش‌های معرفی‌شده در شناسایی مشاهدات مؤثر از طریق داده‌های شبیه‌سازی بررسی می‌شود. داده‌های شبیه‌سازی‌شده شامل دو حالت طراحی متعامد و نامتعامد است.

1. **طراحی متعامد:**
در این حالت، تمامی متغیرهای توضیحی مستقل از یکدیگر در نظر گرفته می‌شوند. با اعمال تغییرات در مقادیر ضرایب، تأثیر مشاهدات مؤثر بر مدل بررسی می‌شود.

2. **طراحی نامتعامد:**
در این حالت، همبستگی بین متغیرهای توضیحی در نظر گرفته می‌شود. نتایج نشان می‌دهند که معیارهای معرفی‌شده به‌طور مؤثری قادر به شناسایی مشاهدات مؤثر هستند.

جدول 5-1:مقادیر تنکی و تفاوت مدل‌ها در طراحی متعامد.

تعداد مشاهدات مؤثر شناسایی‌شده	مقدارb	مقدارa
90	10	5
95	10	10

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

$$Y = 1X_1 + 2X_2 + 3X_3 + 4X_4 + 5X_5 + \varepsilon \qquad (۹)$$

با متغیر پاسخ مرتبط هستند که ε یک نمونه تصادفی با توزیع $N(0,1)$ است؛ ماتریس متغیر کمکی

۱۰۰۰ × ۵۰ به صورت نرمال چندمتغیره با میانگین ۰ و واریانس ۱ تولید می شود و همبستگی دوبه دو بین X_i و X_j برابر است با $\rho^{|i-j|}$. این داده های نرمال شبیه سازی شده مربوط به وضعیتی است که در آن تمام مفروضات رگرسیون OLS استاندارد برقرار است و می توان انتظار داشت که حذف مشاهدات فردی، کمترین تأثیر را بر روی راه حل لاسو داشته باشد. حال به شبیه سازی ها پرداخته می شود تا توانایی معیارهای معرفی شده جهت شناسایی مشاهدات موثر ارزیابی شود. ابتدا مهم است تا بهتر شناخته شوند انواع مشاهداتی که می توانند منجر به تغییراتی در مدل انتخاب شده توسط لاسو گردند. علاوه بر آن ضروری است تا مشخص شود نوع مشاهداتی که می توانند منجر به بی ثباتی مدل یا تغییراتی در تنکی راه حل لاسو شوند. این مشخص سازی با قرارگیری مشاهدات بالقوه موثر در مثال های شبیه سازی زیر کمک می کند. مراحل زیر جهت تولید داده ها برای مطالعه شبیه سازی دنبال شدند:

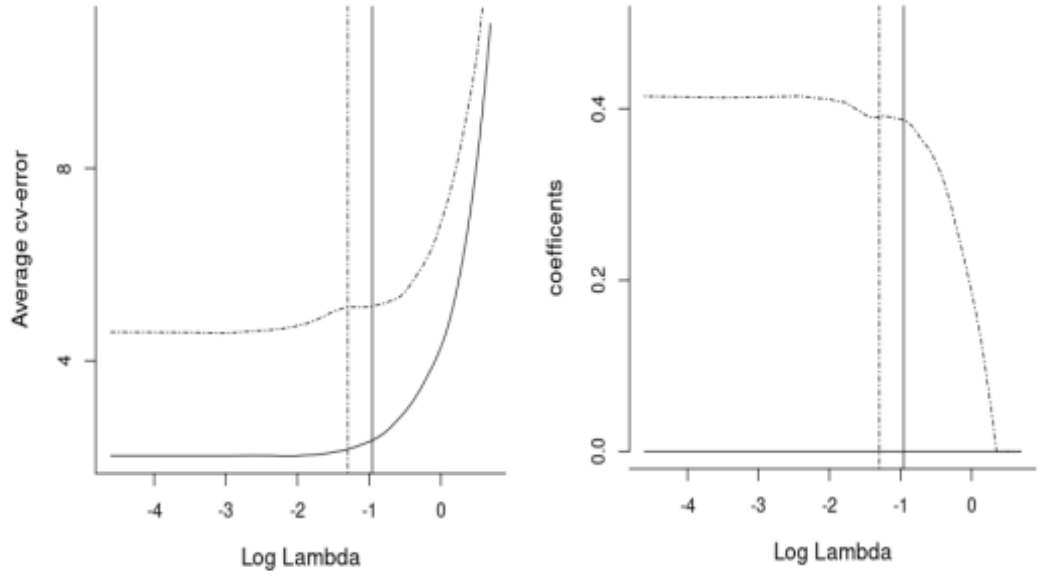
۱- ماتریس متغیرهای کمکی ۱۰۰۰ × ۵۰ از توزیع نرمال چندمتغیره با میانگین ۰ و واریانس ۱ تولید کنید و همبستگی دو به دو بین X_i و X_j برابر است با $\rho^{|i-j|}$. هر دو طرح متعامد $(\rho=0)$ و نامتعامد $(\rho=0.5)$ در نظر گرفته خواهد شد.

۲- از مرحله ۱، ۱۰۰-امین عنصر مشاهده ۱ $\left(x_{1,100}\right)$ را با مقدار a جایگزین کنید.

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

۴- از مرحله ۳، مقدار پاسخ شبیه سازی شده برای مشاهده ۱ (y_1) را با $\mu(y_1)+b$ جایگزین کنید که در آن $\mu(y_1)$ مقدار میانگین y_1 داده شده توسط معادله (۹) است.

تنظیم $|a|$ یا $|b|$ به اندازه کافی بزرگ اجازه می دهد یک مجموعه داده تولید شود که مشاهده ۱ متغیر کمکی یا مقدار(های) پاسخ بزرگ (در قدر مطلق) غیرعادی دارد. این نوع مشاهده انتظار می رود که تأثیر زیادی بر انتخاب مدل لاسو داشته باشد. به طور مشخص از آنجایی که مقدار $x_{1,100}$ را مختل می شود، انتظار می رود (ممکن است) که تغییری در تنکی برای $\hat{\beta}_{100}^{lasso}$ در میان اثرات دیگر به ازای $|a|$ و $|b|$ بزرگ مشاهده شود.



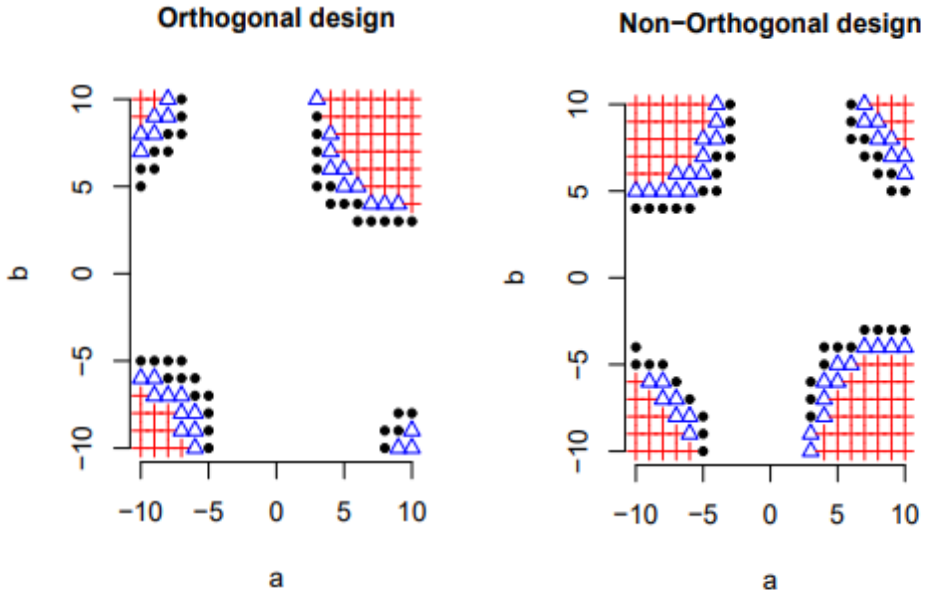
شکل ۱. نمودار خطای اعتبارسنجی متقابل (چپ) و مسیر منظم سازی (راست) $\hat{\beta}_{100}^{lasso}$ از لاسو برازش شده به داده های طراحی نامتعامد شبیه سازی شده با $a=b=$ ۱۰ در حضور و عدم حضور مشاهده ۱

به عنوان مقدمه ای برای شبیه سازی ها نشان داده خواهد شد که چگونه یک مشاهده موثر منحصربه فرد می تواند بر جنبه های خاصی از راه حل لاسو تأثیر بگذارد. یک مجموعه داده طراحی نامتعامد با $a=b=$ ۱۰ تولید و بررسی می شود که چگونه خطای اعتبارسنجی متقابل و مسیر منظم سازی هنگامی که مشاهده موثر (مشاهده ۱) حذف شده، تغییر می کنند. شکل ۱ شامل نمودار خطای اعتبارسنجی متقابل و مسیر منظم سازی برای $\hat{\beta}_{100}^{lasso}$ از لاسو برازش شده به مجموعه داده شبیه سازی شده با و بدون مشاهده ۱ در مدل می شود. 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 ملاحظه ای در خطاهای اعتبارسنجی متقابل و افزایشی در مقدار بهینه λ را نتیجه می دهد.

مسیر منظم سازی $\hat{\beta}_{100}^{lasso}$ برای لاسو کامل در مقابل لاسو برازش شده بدون مشاهده ۱ نیز به طور اساسی متفاوت است. این مسیر برای $\hat{\beta}_{100}^{lasso}$ زمانی که مشاهده ۱ حذف شده در صفر صاف است یعنی پیشگوی X_{100} هرگز وارد مدل نمی شود. با این حال هنگامی که مشاهده ۱ شامل شود، مسیر برای $\hat{\beta}_{100}^{lasso}$ تا زمانی که $\log(\lambda)$ نزدیک صفر نباشد به صفر نمی رسد. از شکل ۱ مشخص است که $\hat{\beta}_{100}^{lasso}>0$ و $\hat{\beta}_{100}^{lasso}(1)=0$. این مثال نشان می دهد که مشاهدات موثر می توانند بر خطای اعتبارسنجی متقابل، مسیر منظم سازی، مقدار بهینه λ و مدل انتخاب شده تأثیر بگذارند. در حقیقت، یک نقطه موثر می تواند به طور بالقوه بر همه چهار معیار تأثیر بگذارد.

مهم است برای طراحی شبیه سازی ها مقادیر a و b پیدا شوند که احتمالاً منجر به تغییر در تنکی برای $\hat{\beta}_{100}^{lasso}$ شدند. برای بررسی این سوال، مجموعه داده های طراحی متعامد و نامتعامد تولید شدند. سپس مقادیر a و b بررسی شدند که منتهی به تغییر در تنکی برای $\hat{\beta}_{100}^{lasso}$ هنگام حذف مشاهده ۱شدند. به ویژه با استفاده از مرحله ۱ طرح تولید داده، داده های متغیر کمکی تولید شدند. هر بار با توجه به اینکه آیا حذف مشاهده ۱ سبب تغییر در تنکی برای $\hat{\beta}_{100}^{lasso}$ شد، مراحل ۲ -۴ با استفاده از داده های تولید شده روی شبکه ای 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 می شود. به بیان دقیق تر برای هر جفت مقادیر a و b روی یک شبکه، نسبت دفعاتی که $\hat{\beta}_{100}^{lasso}$ تغییر در تُنُکی دارد به تصویر کشیده می شود. نمادهای + ، Δ و • به ترتیب مربوط به $p>$ ۰,۷۵ ،

$0,75\geq p>0,5$ و $0,5\geq p>0,25$ هستند. این شکل ها نشان می دهند که تغییرات در تنکی در نواحی رخ می دهند که تقریباً با $|a|\geq 5$ و $|b|\geq 5$ تعریف شده اند به طوری که با افزایش فراوانی تغییر در تنکی، $|a|$ یا $|b|$ از لحاظ اندازه افزایش می یابند. این نتایج به ناپایداری مدل برای داده های تولید شده با $|a|\geq 5$ و $|b|\geq 5$ اشاره می کند. در چنین مواقعی از معیارهای تأثیر معرفی شده انتظار می رود شروع به علامت گذاری مشاهده ۱ کنند.



شکل ۲. نسبت دفعات تغییر در تنکی برای $\hat{\beta}^{lasso}_{100}(p)$ به ازای یک ترکیب معین از a و b به هنگام حذف مشاهده ۱

داده ها برای بررسی اثربخشی معیارهای تأثیر معرفی شده با استفاده از محدوده ای از مقادیر برای a و b تولید می شوند. 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 علامت گذاری کردند یا خیر. همچنین، تعداد مشاهدات نرمال یا غیرموثر علامت گذاری شده ثبت شدند. از مقادیر بُرینش ۲± حاصل شده از لحاظ نظری استفاده می شود برای تعیین اینکه آیا مشاهده ای علامت گذاری می شود یا خیر. نتایج شبیه سازی در جدول ۱ برای محیط نامتعامد آورده می شود.

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

شبیه سازی ها نشان می دهند که معیارهای معرفی شده در علامت گذاری مشاهده ۱ در مواقعی که موثر است، تأثیرگذار هستند. لازم به ذکر است در جدول ۱ "کل" نسبت دفعاتی را نشان می دهد که حداقل یکی از چهار معیار معرفی شده مشاهده ۱ را علامت گذاری 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 علامت گذاری کرد. این ناحیه تأثیر بود که در شکل ۲ نشان داده شد. به همین ترتیب، معیارهای معرفی شده این مشاهده را در مواقعی که $|a| = 2$ و $|b| = 2$ با درصد دفعات بسیار کوچک تری علامت گذاری کردند. این قابل انتظار است زیرا داده ها با واریانس ۱ تولید می شوند پس مقادیر $|a| = 2$ و $|b| = 2$ مربوط به مشاهده ۱ به طور خفیف تأثیرگذار است. Df-model مشاهده مندرج را زمانی که $a = 10$ و $b = 10$ ، ٪۸۱ از مواقع، $a = 5$ و $b = 5$ ، ٪۶۰ درصد از مواقع و $a = 2$ و $b = 2$ ، ٪۸ از مواقع علامت گذاری می کند. به طور کلی، این مقادیر با درصدهای تغییر در تنکی مشاهده شده در شکل ۲ مطابقت دارند و شواهدی ارائه می دهند که df-model مشاهده ۱ را به طور مناسب علامت گذاری می کند.

همچنین، نتایج شکل ۲ نشان می دهد که df-cvpath و df-regpath مشاهده ۱ را با درصد دفعات مناسبی علامت گذاری می کنند. معیار df-regpath تمایل دارد تا مشاهده ۱ را با درصد دفعات بزرگ تری نسبت به df-model علامت گذاری کند. این قابل انتظار است چون که df-regpath معیار حساس تری 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 می دهد. Df-cvpath مشاهده مندرج را هنگامی که $|b| \geq$ ۵، نزدیک به ٪۱۰۰ مواقع علامت گذاری می کند؛ شرایطی که انتظار می رود 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 علامت گذاری می کند. رفتار df-lambda احتمالاً بازتابی از تنوع ذاتی در انتخاب پارامتر منظم سازی برای رگرسیون لاسو است.

مهم تر از همه، شبیه سازی ها نشان می دهند که معیارهای تأثیر معرفی شده برای مشاهدات غیرموثر به درستی کار می کنند به این ترتیب که از این مشاهدات به طور متوسط تقریباً ٪۵ درصد یا کمتر علامت گذاری می شوند. 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 ($b = 0$)، معیارهای معرفی شده به ندرت مشاهده ۱ را علامت گذاری می کنند.

نسبت دفعات تشخیص مشاهده موثر توسط معیارهای تأثیر تحت روش الاستیک نت به ازای α های برابر ۰,۷۵ ،۰,۵۰ و ۰,۲۵ جهت انجام مقایسه در جدول های ۲، ۳ و ۴ برای محیط نامتعامد ارائه می شوند. نتایج شبیه سازی ها نشان می دهند که معیارهای تأثیر تحت الاستیک‌نت تأثیرگذار هستند در علامت گذاری 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

فصل سوم:

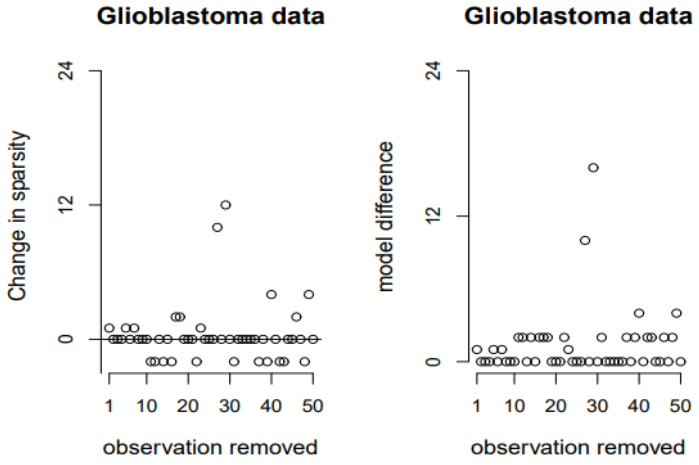
داده‌های واقعی

۶ داده های توصیف ژن گلیوبلاستما

مجموعه داده از مطالعه توصیف ژن ریزآرایه گلیوبلاستما حاصل شده است. میانگین زمان زنده ماندن بیماران مبتلا به گلیوبلاستما، شایع ترین تومور بدخیم مغزی اوّلیه در میان بزرگسالان، ۱۵ ماه است.

هُروث و همکاران (۲۰۰۶) داده های توصیف ژن جهانی را برای ۳۶۰۰ ژن از ۱۲۰ بیمار گلیوبلاستما به دست آوردند (گروه۱= ۵۵ بیمار و گروه ۲= ۶۵ بیمار). در این مقاله، تجزیه و تحلیل به داده های مربوط به گروه ۱ محدود می شود. مشابه تحلیل این داده ها 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 و تفاوت مدل برای بررسی تأثیر روی راه حل لاسو به هنگام حذف مشاهدات فردی مشاهده می شوند. تعداد برآوردهای ضریب صفر، تنکی نامیده می شود. تفاوت مدل، تعداد تفاوت ها در شمول و عدم شمول متغیرهای کمکی بین راه حل لاسو بر اساس مجموعه داده کامل و راه حل لاسو با مشاهده خاص حذف شده، تعریف می شود. به عنوان مثال، اگر حذف یک مشاهده منجر به تغییر متغیرهای کمکی موجود در مدل انتخاب شده از X_3 ،X_4 و X_5 به X_1 ،X_2 و X_3 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 می دهد. این شکل به وضوح نشان می دهد که عدم وجود مشاهدات فردی می تواند به معنای تغییر معنی داری در راه حل لاسو باشد. به عنوان مثال، اگر مشاهده ۲۷ یا ۲۹ در داده های گلیوبلاستما وجود نداشت، تغییر در تنکی به ترتیب برابر ۱۰ و ۱۲ مشاهده می شد. بنابراین، تأثیرات چنین تغییر بزرگی در مدل انتخاب شده توسط لاسو به طور بالقوه بسیار معنی دار است. به ویژه، اگر نتایج به معنای واقعی کلمه در نظر گرفته شوند، می تواند منجر به بررسی بعدی این ۱۰ یا ۱۲ ژن بر اساس داده های تازه به دست آمده شود (در مقایسه با یافته ای که تحقیقات بیشتر بی دلیل است).



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

مطابق شکل ۳ به هنگام حذف مشاهده ۲۷ یا ۲۹ تفاوت مدل به ترتیب برابر ۱۰ و ۱۶ است. لازم به ذکر است که ویژگی های مجموعه داده مورد بررسی در اینجا برای برجسته کردن حساسیت بالقوه لاسو به مشاهدات فردی انتخاب شدند. این مثال به وضوح تأکید می کند بر نیاز به معیارهایی برای سنجش مشاهداتی که پتانسیل اعمال تأثیر قوی بر راه حل لاسو را دارند.

جدول 6-1:ژن‌های برتر شناسایی‌شده بر اساس شمول احتمال.

ژن	شمول احتمال در لاسو	شمول احتمال در الاستیک نت
KCNC1	0.81	0.90
PTEN	0.78	0.86
CNN3	0.76	0.85

با بیش از ۱۰۰ تخصیص تصادفی مشاهدات به دسته ها، تعداد متغیرهای کمکی برای لاسو (الاستیک‌نت با ۱= α) و الاستیک‌نت با α های 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. می توان نتیجه گرفت مدل انتخاب شده پس از حذف مشاهدات موثر و برازش الاستیک‌نت تحت تأثیر تخصیص مشاهدات به دسته های داده ها در اعتبارسنجی متقابل قرار دارد. لذا، اعمال الاستیک‌نت بر روی چندین تخصیص متفاوت مشاهدات به دسته ها (m) به عنوان تحلیلی برای داده های گلیوبلاستما است. نسبت دفعاتی که هر ژن در m مدل انتخاب شده قرار می گیرد، احتمال شمول نامیده می شود که اهمیت نسبی هر ژن را نشان می دهد . ژن هایی با احتمال شمول ۱۰-برتر تحت الاستیک‌نت با α های برابر ۱، ۰,۷۵، ۰,۵۰ و ۰,۲۵ پس از حذف مشاهدات موثر روی ۱۰۰= m تخصیص متفاوت مشاهدات به دسته ها در جدول ۵ ارائه می شود. در خروجی جدول ۵ ژن هایی با احتمال شمول بالا برای الاستیک‌نت با ۱= α (لاسو) نظیر ژن هایی با احتمال شمول بالا برای الاستیک‌نت با ۰,۷۵= α 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 ۰,۷۵= α قرار داشت.

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

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

احتمالات شمول برای همه ژن های ۱۰-برتر تحت لاسو بیشتر از ٪۵۰ هستند که می تواند نشان دهنده ارتباط هر یک از ژن ها با زنده ماندن بیمار باشد. متقابلاً ، تحت الاستیک‌نت با α =۰,۷۵ همه ژن های ۱۰-برتر، احتمالات شمول بیشتر از ٪۵۰ دارند. تفاوت های جزئی بین احتمالات شمول لاسو و الاستیک‌نت با α =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 ژن ها با بالاترین احتمالات شمول لاسو نشان می دهد 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 سرطان های 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 ژن های SLC31A2 و CEBPD احتمالات شمول بزرگی تحت الاستیک‌نت با α =۰,۵۰ و α =۰,۲۵ نسبت به لاسو و الاستیک‌نت با α =۰,۷۵ دارند که بیانگر عدم شناسایی و حذف مشاهده موثر ۲۷ قبل از برازش الاستیک نت با α های مربوطه است. در طول محاسبه احتمالات شمول نیز می توان مشاهده کرد نسبت دفعاتی که هر مشاهده به عنوان مشاهده موثر علامت گذاری می شود. نسبت دفعات با مقادیر بالا برای یک مشاهده نشان دهنده تأثیر بزرگ آن مشاهده بر مدل برازش شده است. برای داده های گلیوبلاستما توسط لاسو و الاستیک‌نت با α های ۰,۷۵، ۰,۵۰و ۰,۲۵

به ترتیب ۲، ۲ ،۱ و ۱ مورد از ۵۰ مشاهده در بیش از ٪۵۰ مواقع به عنوان مشاهده موثر علامت گذاری شدند. لاسو مشاهدات ۲۷ و ۲۹ را به ترتیب در ٪۹۲ و ٪۱۰۰ مواقع علامت گذاری کرد. همچنین، الاستیک‌نت با α =۰,۷۵ مشاهدات ۲۷ و ۲۹ را به ترتیب در ٪۹۱ و ٪۱۰۰ مواقع علامت گذاری کرد اما فقط مشاهده ۲۹ برای الاستیک نت با α =۰,۵۰ و α =۰,۲۵ در ٪۱۰۰ مواقع علامت

گذاری شد. لذا، الاستیک‌نت با ۰,۷۵= α مشابه لاسو مشاهدات ۲۷ و ۲۹ را در بیش از ٪۵۰ مواقع علامت گذاری کرد. این موضوع نشان می دهد که عاقلانه است این مشاهدات برای ارزیابی دلیل تأثیر آنها بیشتر بررسی شوند. بررسی سطحی از داده ها نشان می دهد که مشاهده ۲۹ با ۷ روز کمترین زمان زنده ماندن را دارد و کوچک ترین مقدار بعدی ۴۳ روز است. علاوه بر این، مشاهده ۲۷ مجموعه ای نسبتاً دور از مقادیر توصیف (یا متغیر کمکی) را دارد. مقادیر متغیر کمکی مقیاس پذیر و متمرکز (در مورد انحراف استاندارد و میانگین) این مشاهده مشخص شدند که چهارمین فاصله اقلیدسی بزرگ از میانگین را دارند.

نتیجه‌گیری و پیشنهادات

بحث و نتیجه گیری

دسترسی گسترده به داده های بعد بالا، استفاده از روش های درستنمایی تاوانیده را رایج ساخته است. در این مقاله، توانمندی معیارهای تأثیر معرفی شده تحت لاسو و الاستیک نت با α های برابر ۰,۷۵، ۰,۵۰ و ۰,۲۵ برای تشخیص مشاهدات موثر در داده های بعد بالا به طور نظری بررسی شده و به صورت عددی ارزیابی شده اند. همان طور که از نتایج شبیه سازی ملاحظه می شود، می توان 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 داده های واقعی عملکرد خوبی از خود نشان می دهند زیرا نتایج تعیین مشاهدات موثر و ژن های مهم پس از حذف مشاهدات موثر شناسایی شده توسط این معیارها و برازش مدل الاستیک‌نت به ویژه با α =۰,۷۵ تفاوت چندانی با خروجی تحلیل تحت لاسو ندارد. بی شک نیاز برای معیارهای تشخیص مشاهدات موثر در محیط بعد بالا مهم تر از تحلیل OLS استاندارد است زیرا: (آ) هر مشاهده ای در تحلیل داده های بعد بالا می تواند 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 $p > n$ به طور چشمگیری افزایش داده می شود.

روش‌های تاوانیده نظیر لاسو و الاستیک نت ابزارهای قدرتمندی برای شناسایی مشاهدات مؤثر در داده‌های با ابعاد بالا هستند. نتایج نشان دادند که معیارهای معرفی‌شده به‌طور مؤثری قادر به شناسایی مشاهدات مؤثر بوده و می‌توانند به بهبود مدل و کشف روابط پنهان در داده‌ها کمک کنند.

پیشنهاد می‌شود که این معیارها در سایر حوزه‌ها مانند زیست‌پزشکی و داده‌های مالی نیز مورد استفاده قرار گیرند.

مراجع

معنوی، م. و روزبه، م. (۱۳۹۹)، روش های تحلیل رگرسیونی برای داده های بعد بالا، مجله اندیشه آماری، **۲۵**(۱)،۶۹_۹۰.

نوری، ن. (۱۴۰۱)، تشخیص و آنالیز داده های موثر برای رگرسیون ابعاد بالا، پایان نامه 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.