
بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

رشته آمار، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

مروری بر روش‌های انتخاب متغیر در مدل‌های رگرسیون با بعد بالا

نگارنده: سیده فائزه میرسالاری

استاد راهنما

دکتر محمد آرشی

استاد مشاور

دکتر مینا نوروزی‌راد

شهریور ۱۳۹۸

شماره:
تاریخ:

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای سیده فائزه میرسالاری با شماره دانشجویی ۹۶۱۴۷۸۴ رشته آمار گرایش آمار ریاضی تحت عنوان مروری بر روش‌های انتخاب متغیر در مدل‌های رگرسیون با بعد بالا که در تاریخ ۱۳۹۸/۰۶/۱۳ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

قبول (با درجه خیلی خوب) ☑ **مردود** ☐

نوع تحقیق: **نظری** ☑ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر محمد آرشی	دانشیار	
۲- استادراهنمای دوم	—	—	—
۳- استاد مشاور	دکتر مینا نوروزی راد	—	
۴- نماینده تحصیلات تکمیلی	دکتر محمدرضا ربیعی	استادیار	
۵- استاد ممتحن اول	دکتر داود شاهسونی	دانشیار	
۶ـ استاد ممتحن دوم	دکتر حسین باغیشنی	استادیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده:

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم خالصانه و خاضعانه به:

پدرم به پاس سال‌ها حمایت‌های بی‌دریغش
مادرم به پاس سال‌ها صبوری‌های بی‌انتهایش

سپاس گزاری

منت خدای را عزوجل که طاعتش موجب قربت است و به شکر اندرش مزید نعمت.

به مصداق لم یشکر المخلوق، لم یشکر الخالق،
سپاس بی‌کران نثار پدر و مادرم، که ناتوان شدند تا من توانا شوم، مویشان سپید شد تا من روسفید شوم و عاشقانه سوختند تا گرمابخش وجودم و روشنگر راهم باشند. اگر پدرم و حمایت‌های بی‌دریغش نبود، این سطرها هیچ‌گاه نگاشته نمی‌شد.

سپاسگزاری و قدردانی خالصانه‌ام نثار اساتیدم جناب آقای دکتر محمد آرشی و سرکار خانم دکتر مینا نوروزی‌راد که با حمایت‌های بی‌دریغشان تاکنون هیچ‌گاه تنهایم نگذاشته‌اند و با صبر و تلاش فراوان ایشان توانستم پایان‌نامه کارشناسی ارشد را گردآوری نمایم.

از اساتید محترم، دکتر احمد نزاکتی، دکتر داود شاهسونی، دکتر حسین باغیشنی، دکتر محمدرضا ربیعی که در دوران تحصیلی‌ام صادقانه و عاشقانه تلاش نمودند تا برایم نقالی دانسته‌ها را به نقادی اندیشه‌ها تبدیل کنند، قدردانی می‌کنم.

از مدیر گروه دلسوز و پرتلاش گروه آمار دانشگاه صنعتی شاهرود، جناب آقای دکتر حسین باغیشنی به دلیل خدمات، راهنمایی‌ها و زحماتشان سپاسگزارم.

از اساتید داور محترم، جناب آقای دکتر داود شاهسونی و دکتر حسین باغیشنی که با حضور دلگرمشان تصحیح و داوری این پایان‌نامه را بر عهده گرفته‌اند سپاسگزارم و دوباره بارها و بارها از همدلی و همراهی و همگامی **خانواده مهربانم** که همواره با قلبی آکنده از عشق و معرفت محیطی سرشار از سلامت، امنیت، آرامش و آسایش برایم فراهم آورده‌اند و مرا در رسیدن به اهدافم یاری رسانده‌اند، بی‌نهایت سپاس‌گزارم.

سیده فائزه میرسالاری
شهریور ۱۳۹۸

تعهد نامه

اینجانب **سیده فائزه میرسالاری** دانشجوی کارشناسی ارشد رشته **آمار** دانشکده **علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **مروری بر روش‌های انتخاب متغیر در مدل‌های رگرسیون با بعد بالا** ، تحت راهنمایی **دکتر محمد آرشی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام '' دانشگاه صنعتی شاهرود '' یا " Shahrood University of Technology " به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

سیده فائزه میرسالاری
شهریور ۱۳۹۸

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

امروزه به دلیل توسعه تکنولوژی، ذخیره و تحلیل داده‌ها با تعداد متغیرهای زیاد امکان‌پذیر شده است. اما باید به این نکته توجه داشت که چنان‌چه متغیرهای ثبت شده معنی‌دار نبوده یا تعداد آن‌ها زیاد باشد، تکنیک‌های برآورد کلاسیک کارا نبوده و علاوه بر شناسایی و حذف متغیرهای زائد باید از روش‌هایی غیرکلاسیک برای برآورد پارامترهای مدل استفاده کرد. در مدل‌بندی رگرسیونی چنان‌چه برآوردگرهای جریمه‌شده را به‌کار ببریم قادریم هم متغیرهای معنی‌دار را انتخاب کرده و هم پارامترها را برآورد کنیم. در این راستا، چنان‌چه متغیرها طوری گروه‌بندی شده باشند که باید تمامی آن‌ها حذف شوند یا باید همگی در مدل باقی بمانند، می‌توان از برآوردگرهای جریمه‌شده گروهی استفاده کرد. در این پایان‌نامه، با مروری مختصر پیرامون برخی روش‌های مرسوم و پرکاربرد انتخاب متغیر مدرن، برآوردگرهای جریمه‌شده گروهی را مورد بررسی قرار داده و با استفاده از یک سری مطالعات عددی، شبیه‌سازی شده و تحلیل داده واقعی ژنومی کاربرد آن‌ها را در مدل رگرسیون لجستیک تحقیق می‌کنیم.

کلمات کلیدی: اسکد، انتخاب متغیر، بعد بالا، داده‌های ژنومی، لاسوی گروهی.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

معرفی و بیان مسئله

در سال‌های اخیر، پیشرفت‌های بزرگی از روش‌های بعد بالا برای آزمایش‌های زیست‌شناسی مولکولی به‌وجود آمده‌اند. ظهور تکنولوژی‌های میکروآرایه [۱]، دانشمندان را به اندازه‌گیری همزمان ده‌ها هزار ژن‌ها قادر ساخته است. تکنولوژی میکروآرایه DNA، به محققان این اجازه را می‌دهد که میان هزاران پرونده ژنی در یک موجود زنده روی یک محیط کشت آزمایشی مجزا، اثرات متقابل را بررسی کنند. پیش از این تکنولوژی، محققان محدود به انجام آزمایشات با تعداد بسیار کوچک‌تری از واحدهای ژنتیکی در هر آزمایش بودند و قادر بودند اثرات متقابل میان ژن‌ها را تحت تغییر شرایط، در روی مقیاس بسیار کوچک‌تری تعیین کنند.

چندریختی نوکلئوتید منفرد[۲] (SNP) مبتنی بر مطالعات ارتباطی[۳] به شناسایی SPN های در ارتباط با فنوتیپ‌ها، برای مثال بیماری‌های پیچیده، کمک می‌کند. SNPها ممکن است با بیماری، خودشان به‌تنهایی به‌عنوان تأثیرات اصلی ارتباط داشته باشند یا در کنار سایر عوامل، به‌عنوان تعامل‌های شناختی رفتار کنند.

برای تحلیل داده‌ها با بعد بالا، مشکل اصلی این است که تعداد متغیرها، برای مثال SPNها یا ژن‌ها، به‌سبب هزینه بالای آزمایش‌های میکروآرایه‌ای به‌مراتب بیش‌تر از تعداد نمونه‌ها است. اگر شناسایی کردن اثرات متقابل نیز جز اهداف مورد نظر محقق باشد، این مشکل تشدید می‌شود. از این رو، باید متغیرهای بی‌اهمیت حذف شوند و یک زیرمجموعه از متغیرها

[1] Microarray
[2] Single Nucleotide Polymorphism (SNP)
[3] Association

که موثرترین اثرها را توضیح می‌دهند، در مدل باقی بمانند. وست و همکاران (۲۰۰۱)، در مقاله‌شان، این مسئله را مسئله «p بزرگ و n کوچک» می‌نامند. هنگامی‌که تعداد نمونه‌های n، معمولا درحدود ده‌ها یا صدها اما تعداد متغیرهای p معمولا در حدود هزارها یا ده‌ها هزار باشد، در برازش مدل‌های رگرسیونی، این مسئله به‌وجود می‌آید.

۱. تعداد پاسخ‌های نامتناهی: اگر $p > n$، آن‌گاه تعداد مجهول‌ها بیش‌تر از معادله‌ها خواهد بود و ممکن است جواب‌های بی‌شماری وجود داشته باشند.

۲. بیش برازشی مدل: مدل، داده‌های مدل‌ساز [۴] را به خوبی برازش می‌دهد اما در مورد داده‌های آزمون، به‌خوبی رفتار نمی‌کند.

۳. همخطی چندگانه: ژن‌های زیادی الگوهای تقریبا مشابهی را در میان نمونه‌ها نشان خواهند داد، بنابراین آن‌ها اطلاعات جدیدی را عرضه نمی‌کنند؛ خصوصیات بعضی ژن‌ها می‌تواند ترکیب خطی از خصوصیات ژن‌های دیگر باشد.

روش‌های انتخاب متغیر به‌طور گسترده‌ای در مطالعات بررسی شده‌اند. در این زمینه، می‌توانید جورج و مک کولا (۱۹۹۳)؛ برایمن (۱۹۹۵)؛ تیبشیرانی (۱۹۹۶)؛ جورج و فاستر (۲۰۰۰)؛ فن و لی (۲۰۰۱)؛ شن و یی (۲۰۰۲)؛ افرون و همکاران (۲۰۰۴)؛ ژو و هیستی (۲۰۰۵) و وو و همکاران (۲۰۰۷) را ببینید. به‌ویژه، لاسو (تیبشیرانی، ۱۹۹۶) در چند سال اخیر توجهات زیادی را به‌خود جلب کرده است. مسئله «p بزرگ و n کوچک» را می‌توان با اضافه کردن یک جریمه به مدل رگرسیونی کلاسیک، حل کرد. لاسو، مجموع توان‌های دوم باقی‌مانده‌ها را به شرط اینکه مجموع مقدار قدرمطلق ضرایب کمتر از یک مقدار ثابت شود، کمینه می‌کند. به‌دلیل ماهیت این محدودیت (شرط)، تعدادی از ضرایب به اجبار، دقیقا صفر و از این رو، ژن‌ها یا SNPهای نامرتبط در مدل حذف خواهند شد. جریمه نوع L_1 لاسو همچنین می‌تواند به دیگر مدل‌ها، به‌عنوان مثال، رگرسیون لجستیک (لوخارست، ۱۹۹۹؛ روث، ۲۰۰۴؛ شواده و کیتهی، ۲۰۰۳) با جایگزین کردن مجموع توان‌های دوم باقی‌مانده با منفی تابع لگاریتم درستنمایی متناظر اعمال شود. مسئله دیگری که محققان در دنیای کاربردی با آن مواجه هستند، این است که گاهی اوقات، متغیرها به یکدیگر وابسته هستند و از این رو «گروه‌بندی شده» هستند. برای مثال در مدل تحلیل واریانس یک عامل ممکن است چندین سطح داشته باشد و با چند متغیر موهومی [۵] بیان شود. متغیرهای موهومی از یک عامل، یک گروه را تشکیل می‌دهند. در کاربردهای زیستی، ژن‌هایی که یک تابع زیستی مشترک را سهیم می‌شوند یا در مسیر متابولیک یکسانی شرکت می‌کنند، همبستگی جفتی [۶] بالایی با یکدیگر دارند. بنابراین این نوع ژن‌ها یک گروه می‌شوند.

[4] Training data
[5] Dummy
[6] Pairwise correlation

وقتی ژن‌ها گروه‌بندی شده هستند، معمولا خیلی قطعی نیست که ساختار گروه نادیده گرفته شود و رگرسیون جریمه‌شده به‌کار گرفته شود. برای مثال، فرض کنید تاثیر گروه kام بر متغیر پاسخ بی‌اهمیت باشد. بنابراین با استفاده از روش لاسو، به‌جای اینکه کل گروه برابر صفر قرار بگیرد، فقط یک ضریب از گروه kام، صفر می‌شود. بنابراین، لاسو به‌جای گروه، برای متغیرهای مجزا انتخاب متغیر را بر روی یک متغیر اجرا می‌کند که اغلب منجر به انتخاب متغیرهای بی‌اهمیت بیشتری می‌شود.

لاسوی گروهی (باکین، ۱۹۹۹؛ کای، ۲۰۰۱؛ آنتونیادیس و فن، ۲۰۰۱؛ یوان و لین، ۲۰۰۶) با معرفی یک بسط مناسب این مسئله را حل کرده است. یوان و لین (۲۰۰۶) لاسوی گروهی را پیشنهاد کردند که تعمیمی از لاسو برای انجام انتخاب متغیر روی گروه‌های (از پیش تعریف‌شده) متغیرها در مدل‌های رگرسیون خطی است. تابع جریمه از نرم‌های L_2 گروه‌ها تشکیل شده است که باعث می‌شود اثر تُنُکی در سطح گروهی ایجاد شود در حالی‌که جریمه مشابه ریج درون گروه، به‌کار برده می‌شود. لاسو گروهی نیز نقطه ضعف‌های خود را دارد. لاسو گروهی، قادر به انجام انتخاب متغیر در سطح‌های متغیرهای تکی نیست و هم‌چنین ضرایب بزرگ را به‌شدت منقبض می‌کند. مایر و همکاران (۲۰۰۸) این ایده را به‌رگرسیون لجستیک تعمیم دادند.

از روش‌های دیگری که در مقاله‌ها و تحقیقات، معرفی شده‌اند و انتخاب متغیر گروهی را بهبود می‌دهند، می‌توان به بریج (فرانک و فریدمن، ۱۹۹۳)، SCAD (فن و لی، ۲۰۰۱) و MCP (ژانگ، ۲۰۱۰) اشاره کرد. این رهیافت‌ها، انتخاب متغیر را هم‌چنان برای گروه‌ها انجام می‌دهند و قادر نیستند متغیرها را جداگانه خارج از گروه انتخاب کنند.

در راستای پیدا کردن یک راه‌حل برای این مسئله، هوانگ و همکاران (۲۰۰۹) رهیافت بریج گروهی را پیشنهاد کردند که انتخاب متغیر را با ایجاد پاسخ‌های تنک در دو سطح گروه و تک تک متغیرها، اجرا می‌کند. بریهنی و هوانگ (۲۰۰۹) روشی را برای انتخاب گروه‌های مهم تعریف کردند که به‌خوبی شناسایی متغیرهای بااهمیت درون این گروه‌ها رفتار می‌کند. در واقع، یک انتخاب دو-سطحی[۷] انجام می‌دهد. آن‌ها برآوردگر جدیدی به‌نام MCP گروهی معرفی کردند که انتخاب متغیر را در دو سطح گروه و تک تک متغیرها انجام می‌دهد. همجنین آن‌ها برآوردگرهای لاسو گروهی، بریج گروهی و MCP گروهی را بررسی کردند و یک چارچوب جدید برای جریمه‌های گروهی معرفی کردند. در این پایان‌نامه، در فصل اول، برخی از مفاهیم اولیه و نیز تعاریف کلی از برخی روش‌های ارائه شده را آورده‌ایم. روش‌هایی که در بالا برای انتخاب متغیر در رگرسیون بعد بالا مطرح شد، در خانواده روش‌های انتخاب متغیر مدرن قرار می‌گیرند که قبل از بررسی تفصیلی آن‌ها در فصل ۳، در فصل ۲ روش‌های انتخاب متغیر کلاسیک بیان می‌شود. فصل ۴ این پایان‌نامه، به‌روش‌های انتخاب متغیر گروهی اختصاص دارد.

[7] Bi-level selection

۱.۱ انتخاب مدل

یک بردار ورودی حقیقی‌ـ‌مقدار $\mathbf{X} \in \mathbb{R}^p$ (بردار پیشگوها نیز نامیده می‌شود) و یک عدد خروجی حقیقی‌ـ‌مقدار $Y \in \mathbb{R}$ (پاسخ نیز نامیده می‌شود) را درنظر بگیرید. توجه اصلی به مدل‌های آماری محدود شده است که Y به‌صورت

$$Y = f(\mathbf{X}) + \epsilon$$

به $\mathbf{X}$ ارتباط دارد که $\mathrm{E}(\epsilon) = ۰$. این پذیره به‌معنای این است که همه انحراف‌ها از رابطه مشخص $Y = f(\mathbf{X})$ می‌توانند با یک خطای افزودنی ϵ در نظر گرفته شوند.

در مسئله یادگیری آماری یک مجموعه مدل‌ساز[۸] $\mathcal{T}$ از N زوج مرتب $(\mathbf{X}_۱, Y_۱)$، ...، $(\mathbf{X}_N, Y_N)$ در دسترس است که اغلب به‌صورت یک ماتریس و یک بردار در کنار یکدیگر، $\mathcal{T} = (\mathbf{X}, \mathbf{Y})$، گروه‌بندی می‌شوند. هدف، استفاده از این مجموعه مدل‌ساز برای یافتن تابع $\hat{f}(\mathbf{X})$ است که بهترین برآورد $f(\mathbf{X})$ را ارائه می‌دهد که از نماد $\hat{Y}$ برای نمایش آن استفاده می‌شود. به‌عبارت دیگر،

$$\hat{Y} = \hat{f}(\mathbf{X}).$$

یک ویژگی کلیدی از هر مدلی که ممکن است انتخاب شود، پیچیدگی‌اش است. مدل‌های پیچیده‌تر از تعداد زیادی از متغیرها در $\mathbf{X}$ برای برآورد Y استفاده می‌کنند ـ بنابراین، آن‌ها بعد بالا هستند و به پارامترهای زیادی نیاز دارند. در مدل‌هایی با پیچیدگی کمتر، تنها تعداد کمی از متغیرهای $\mathbf{X}$ برای برآورد Y استفاده می‌شوند ـ آن‌ها بعد پایین هستند و پارامترهای کمتری نیاز دارند. اگر $\mathbf{X}$ از تعداد متغیرهای زیادی تشکیل شده باشد، معمولا مدل‌های با بعد پایین منجر به مدل‌های بهتری می‌شوند.

مطئنا، یافتن بهترین برآورد مستلزم به‌کارگیری یک معیار برای میزان خوبی یک برآوردگر است. این معیار، تابع زیان[۹] گفته می‌شود و با $L(Y, \hat{Y})$ نشان داده می‌شود. فرض کنید یک بردار ورودی $\mathbf{X}$ و یک خروجی درست و واقعی[۱۰] Y در دسترس است. تابع زیان بیان می‌کند پیش‌بینی $\hat{Y} = \hat{f}(\mathbf{X})$ چه‌قدر خوب است. رایج‌ترین و مناسب‌ترین تابع زیان، زیان توان‌های دوم خطا [۱۱] است.

تعریف ۱.۱.۱ (زیان توان‌های دوم خطا)**.**

$$L(Y, \hat{Y}) = (Y - \hat{Y})^۲.$$

[8] Training set
[9] Loss function
[10] True
[11] Squared error loss

از دیدگاه این تابع زیان، هدف یافتن $\hat{f}(\mathbf{X})$ است که خطای کلی مورد انتظار [۱۲] را کمینه می‌کند.

تعریف ۲.۱.۱ (خطای کلی مورد انتظار).

$$\mathrm{Err} = \mathrm{E}_{(x,y)}\{L(Y,\hat{Y})\}.$$

در این پایان‌نامه، انواع مختلفی از امیدهای ریاضی E با توجه به فضای احتمال تعریف می‌شوند.

تعریف ۳.۱.۱ (امید ریاضی یا مقدار مورد انتظار). انواع مختلفی از امیدهای ریاضی که در این پایان‌نامه استفاده شده، به‌صورت‌های زیر تعریف می‌شوند:

- $\mathrm{E}_{(x,y)}$ نشان می‌دهد که میانگین روی هر زوج‌مرتب از ورودی‌ها و خروجی‌ها گرفته می‌شود.
- $\mathrm{E}_{(X,Y)\in\mathcal{T}}$ بیان می‌کند که میانگین روی همه زوج مرتب‌های (X,Y) که فقط در مجموعه مدل‌ساز $\mathcal{T}$ محاسبه می‌شود.
- $\mathrm{E}_{\mathcal{T}}$ نشان می‌دهد که میانگین روی تمام مجموعه‌های مدل‌ساز گرفته می‌شود.
- $\mathrm{E}_{Y^{\mathrm{NEW}}}$ برای تعریف خطای درون‌نمونه‌ای [۱۳] استفاده می‌شود، که در ادامه بیان می‌شود اینجا، Y^{NEW} یک مشاهده جدید Y به‌ازای ورودی داده‌شده $\mathbf{X}$ نشان می‌دهد. این امید ریاضی بیان می‌کند که میانگین روی تمام این مشاهده‌های جدید گرفته شده است.

در استفاده از تعریف ۲.۱.۱، یک مشکل مطرح است. تنها N مشاهده (آن‌هایی که در $\mathcal{T}$ هستند) برای آزمون مدل پیشنهادی وجود دارد، اما باید خطای کلی مورد انتظار بر اساس همه مشاهده‌های ممکن کمینه شود.
معمولا پاسخ به این مسئله، تنها مورد توجه قرار دادن داده‌های مجموعه مدل‌ساز است. به‌عبارت دیگر، به‌جای کمینه کردن خطای کلی مورد انتظار، خطای مدل‌ساز [۱۴] کمینه شود.

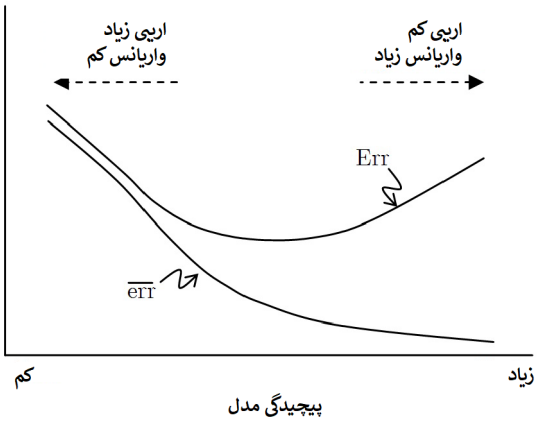
تعریف ۴.۱.۱ (خطای مدل‌ساز).

$$\begin{aligned}\overline{\mathrm{err}} &= \mathrm{E}_{(X,Y)\in\mathcal{T}}\{L(Y,\hat{Y})\} \\ &= \frac{۱}{N}\sum_{(X,Y)\in\mathcal{T}} L(y,\hat{y})\end{aligned}$$

[12] Expected generalization error
[13] In-sample error
[14] Training error

مسئله

برای مسائل بعد پایین (که در آن $\mathbf{x}$ یک بردار کوچک است)، خطای مدل‌ساز یک برآورد نسبتا خوب برای خطای کلی مورد انتظار است. اگرچه، هرچه مدل‌های بزرگ‌تری درنظر گرفته شود، شرایط به‌سرعت بدتر می‌شود. شکل ۱.۱ دو مسئله حل شده را نشان می‌دهد.



شکل ۱.۱: رفتار خطای مدل‌ساز و خطای کلی موردانتظار نسبت به تغییرات پیچدگی مدل.رفتار خطای مدل‌ساز و خطای کلی موردانتظار نسبت به تغییرات پیچدگی مدل. خطای مدل‌ساز واریانس مدل را به‌حساب نمی‌آورد و از این رو همیشه خطای کلی موردانتظار را به‌ویژه برای مدل‌های بسیار پیچیده، کمتر برآورد می‌کند. این شکل از هیستی و همکاران (۲۰۰۹) اقتباس شده است.

- خطای مدل‌سازی همیشه خطای کلی موردانتظار را کم‌تر برآورد می‌کند.
- به‌طور نگران‌کننده ای، رفتارهای دو کمیت برای مدل‌های پیچیده‌تر متفاوت هستند؛ خطای مدل‌ساز با افزایش پیچیدگی، کاهش می‌یابد، درحالی‌که خطای کلی موردانتظار به کمترین مقدار خود می‌رسد. بدین معنی که بهترین مدل در سطح متوسطی از پیچیدگی قرار می‌گیرد، یک کمینه‌سازی ساده خطای مدل‌ساز منجر به انتخاب پیچیده‌ترین مدل ممکن می‌شود.

هر دو مورد بیان شده را می‌توان با تعریف خوش‌بینی موردانتظار[۱۵] به‌صورت یک مقدار کمی بیان کرد که بیان می‌کند خطای مدل‌ساز، خطای کلی موردانتظار را کمتر برآورد می‌کند.

تعریف ۵.۱.۱ (خوش‌بینی موردانتظار).

$$\omega = \mathrm{E}_{\mathcal{T}}\{\mathrm{Err} - \overline{\mathrm{err}}\}$$

که همان‌طور که اشاره شد، امیدریاضی روی همه مجموعه‌های مدل‌ساز ممکن $\mathcal{T}$ گرفته می‌شود.

[15] Expected optimism

در ادامه، دو مورد مطرح شده با جزئیات بیشتر مورد توجه قرار می‌گیرند.

اولین مسئله به‌طور شهودی قابل درک است. با توجه به اینکه مجموعه مدل‌ساز $\mathcal{T}$ برای برازش مدل $\hat{f}(\mathbf{X})$ استفاده شده است، مدل برای داده‌های مجموعه مدل‌ساز متناسب است. از این رو، مدل برای مشاهدات در مجموعه $\mathcal{T}$ ($\overline{\text{err}}$ پایین) بسیار بهتر از هنگامی‌است که از مشاهداتی که درون مجموعه $\mathcal{T}$ نیستند ($\overline{\text{Err}}$ بالا) عمل می‌کند. خطای مدل‌ساز از داده‌های $\mathcal{T}$ برای برازش مدل استفاده می‌کند که نیکویی برازش را بیان می‌کند. خطای کلی موردانتظار از یک داده جدید استفاده می‌کند که برازش را بیان می‌کند.

برای درک دومین مسئله، باید ماهیت خطای کلی موردانتظار با جزئیات بیشتری بررسی شود.

قضیه ۱.۱.۱. خطای کلی موردانتظار با در نظر گرفتن تابع زیان توان دوم خطا می‌تواند به‌صورت رابطه (۱.۱) نوشته شود

$$\mathrm{E}_{(x,y)}\left\{L(y,\hat{f}(\mathbf{x}))\right\} = \mathrm{V}(\epsilon) + \mathrm{V}(\hat{f}(\mathbf{x})) + \left[\mathrm{E}\left(\hat{f}(\mathbf{x})\right) - f(\mathbf{x})\right]^{2} \qquad (۱.۱)$$

برهان. بر اساس تعریف ۱.۱.۱، می‌توان نوشت

$$\begin{aligned}
\mathrm{E}_{(x,y)}\left[L(y,\hat{f}(\mathbf{x}))\right] &= \mathrm{E}\left[\left(y-\hat{f}(\mathbf{x})\right)^{2}\right] \\
&= \mathrm{E}\left[\left(f(\mathbf{x})+\varepsilon-\hat{f}(\mathbf{x})\right)^{2}\right] \\
&= \mathrm{E}\left[\epsilon^{2}+\left(f(\mathbf{x})-\hat{f}(\mathbf{x})\right)^{2}+2\epsilon\left(f(\mathbf{x})-\hat{f}(\mathbf{x})\right)\right]
\end{aligned}$$

با توجه به اینکه $\mathrm{E}[\epsilon] = 0$ و ϵ مستقل از $\mathbf{X}$ است، $\mathrm{E}\left[\epsilon\left(f(\mathbf{x})-\hat{f}(\mathbf{x})\right)\right] = 0$. هم‌چنین، $\mathrm{E}\left[\epsilon^{2}\right] = \mathrm{V}(\epsilon)$. بنابراین،

$$\begin{aligned}
\mathrm{E}\left[L(y,\hat{f}(\mathbf{x}))\right] &= \mathrm{V}(\epsilon) + \mathrm{E}\left[\left(f(\mathbf{x})-\hat{f}(\mathbf{x})\right)^{2}\right] \\
&= \mathrm{V}(\epsilon) + \mathrm{V}(\hat{f}(\mathbf{x})) + \left[\mathrm{E}\left[\hat{f}(\mathbf{x})\right] - f(\mathbf{x})\right]^{2}
\end{aligned}$$

و اثبات کامل است. □

سه عبارت رابطه (۱.۱) را درنظر بگیرید.

- $\mathrm{V}(\epsilon)$ خطای ساده‌نشدنی[16] در Y نامیده می‌شود ـ که از مدل واقعی و مورد نظر f می‌آید. حتی اگر مدل آماری $\hat{f}$ دقیقا مدل موردنظر واقعی را نشان دهد، هم‌چنان این خطا را در بر دارد.
- $\mathrm{V}(\hat{f}(\mathbf{x}))$ واریانس مدل $\hat{f}$ است که یک منبع تصادفی[17] خطا در این مدل می‌باشد.

[16] Irreducible error
[17] Random

- $\left[\mathrm{E}\left[\hat{f}(\mathbf{x})\right] - f(\mathbf{x})\right]^2$ اریبی[18] مدل $\hat{f}$ است. این یک منبع سیستماتیک [19] از خطا در مدل است.

چگونه اریبی و واریانس با تغییر پیچیدگی مدل $\hat{f}$، تغییر می‌کنند؟

- برازش یک مدل با تعداد متغیرهای زیاد منجر به یک مدل با اریبی کم می‌شود. این موضوع، به این دلیل اتفاق می‌افتد که مدل‌های پیچیده‌تر، داده‌ها را بهتر برازش خواهند داد. همچنین بدین معنی است که پارامترهای بیشتری با تعداد داده‌های یکسان برای برآورد وجود دارند و به‌عبارت دیگر، واریانس هر پارامتر بسیار بزرگ خواهد بود.
- برازش یک مدل ساده با تعداد متغیرهای کم منجر به مدلی با واریانس کم می‌شود، زیرا تعداد زیادی مشاهدات برای برآورد هر پارامتر در دسترس خواهند بود. با این حال، از آن‌جایی‌که پیچیدگی مدل کمتر خواهد بود، اریبی بزرگتری نتیجه می‌شود.

این رفتار خطای کلی موردانتظار، موازنه اریبی‌ـ‌واریانس[20] نامیده می‌شود. خطای مدل‌ساز هیچ یک از این جزئیات را نشان نمی‌دهد. تنها اریبی را مورد توجه قرار می‌دهد و به واریانس توجهی ندارد. هرچه مدل پیچیده‌تری انتخاب شود، به همان اندازه داده‌ها در مجموعه مدل‌ساز بهتر برازش می‌شوند و منجر به خطای مدل‌ساز کمتری می‌شود. تفاوت رفتارهای $\overline{\mathrm{err}}$ و Err در شکل ۱.۱ به‌خوبی توضیح داده شده است.

می‌توان با استفاده از قضیه ۲.۱.۱ نگاه دقیق‌تری به موارد مطرح‌شده داشت. ابتدا، با تغییر اندکی در تعریف خطای کلی موردانتظار، خطای درون‌ـ‌نمونه‌ای به‌صورت تعریف ۶.۱.۱ تعریف می‌شود.

تعریف ۶.۱.۱ (خطای درون‌ـ‌نمونه‌ای).

$$\mathrm{Err}_{\mathrm{in}} = \mathrm{E}_{Y^{\mathrm{New}}}\mathrm{E}_{(\mathbf{X},Y)\in\mathcal{T}}\left[L(Y^{\mathrm{NEW}}, \hat{f}(\mathbf{X}))\right] = \frac{1}{N}\sum_{i=1}^{N} L(\bar{Y}_i, \hat{Y}_i)$$

که Y^{NEW} یک مشاهده جدید در نقطه $\mathbf{X}_i$ را بیان می‌کند، $\bar{Y}_i = \mathrm{E}\left[f(\mathbf{X}_i)\right]$ میانگین همه مشاهدات است و $\hat{Y}_i = \hat{f}(\mathbf{X}_i)$.

شباهت این خطا با خطای کلی موردانتظار (تعریف ۲.۱.۱) بدیهی است. هر دو خطا روی نقاط داده جدید بیرون از مجموعه مدل‌ساز تعریف می‌شوند. تنها تفاوت این است که خطای درون‌ـ‌نمونه‌ای این داده جدید را محدود می‌کند به‌طوری‌که باید نقاط داده جدید در مختصات $\mathbf{X}$ یکسان با آن نقطه در مجموعه مدل‌ساز باشند.

بنابراین تعریف جدیدی از خوش‌بینی موردانتظار به‌صورت زیر ارائه می‌شود.

[18] Bias
[19] Systematic
[20] Bias-Variance tradeoff

تعریف ۷.۱.۱ (خوش‌بینی موردانتظار (بازنویسی‌شده)).

$$\omega = \mathrm{E}_{\mathcal{T}}\left[\mathrm{Err}_{\mathrm{in}} - \overline{\mathrm{err}}\right]$$

در ادامه برای اثبات قضیه ۲.۱.۱ لم زیر را بیان می‌کنیم:

لم ۱.۱.۱. فرض کنید ℓ یک تابع مقعر باشد و خانواده‌ای از توابع زیان به‌صورت

$$L(y_i, \hat{y}_i) = \ell(\hat{y}_i) + \dot{\ell}(\hat{y}_i)(y_i - \hat{y}_i) - \ell(y) \qquad (۲.۱)$$

که در آن $\dot{\ell}$ به مفهوم مشتق تابع ℓ است، در دسترس باشند. برای توابع زیان در این خانواده،

$$\omega = \frac{۱}{N}\sum_{i=۱}^{N} \mathrm{E}_{\mathcal{T}}\left[\hat{\psi}_i(y_i - \bar{y}_i)\right]$$

که در آن ψ برداری است که iامین مولفه آن $\hat{\psi}_i \equiv -\dot{\ell}(\hat{y}_i)$ است.

ω خوش‌بینی موردانتظار (تعریف ۷.۱.۱)، N تعداد اعضای مجموعه مدل‌ساز $\mathcal{T}$، y_i، iامین خروجی در مجموعه مدل‌ساز، $\bar{y}_i$ میانگین همه خروجی‌های Y_i برای ورودی $\mathbf{X}_i$ و هم‌چنین $\hat{Y}_i = \hat{f}(\mathbf{X}_i)$، پیش‌بینی مدل از آن‌چه Y_i باید باشد.

برهان. برای توابع زیانی که در رابطه (۲.۱) صدق می‌کنند می‌توان نوشت:

$$\begin{aligned} L(y_i^{\mathrm{NEW}}, \hat{y}_i) - L(y_i, \hat{y}_i) &= -\dot{\ell}(\hat{y})\left(y_i - y_i^{\mathrm{NEW}}\right) + \ell(y_i) - \ell(y_i^{\mathrm{NEW}}) \\ &= \hat{\psi}_i(y_i - y_i^{\mathrm{NEW}}) + \ell(y_i) - \ell(y_i^{\mathrm{NEW}}). \end{aligned}$$

بنابراین

$$\mathrm{Err}_{\mathrm{in}} - \overline{\mathrm{err}} = \frac{۱}{N}\sum_{i=۱}^{N} \hat{\psi}_i(y_i - \bar{y}_i) + \ell(y_i) - \ell(y_i^{\mathrm{NEW}}). \qquad (۳.۱)$$

در نهایت، از آن‌جایی که $\omega = \mathrm{E}_{\mathcal{T}}\left[\mathrm{Err}_{\mathrm{in}} - \overline{\mathrm{err}}\right]$، هنگامی که امید ریاضی نسبت به همه مجموعه‌های مدل‌ساز گرفته می‌شود، دو عبارت آخر در رابطه (۳.۱) برابر صفر است زیرا $\mathrm{E}\left[y_i^{\mathrm{NEW}}\right] = y_i$ و بنابراین، $\mathrm{E}\left[\ell(y_i^{\mathrm{NEW}})\right] = \ell(y_i)$. با توجه به ویژگی خطی بودن امید ریاضی، می‌توان جای امید ریاضی و نماد $\sum$ را جا به جا کرد و در نتیجه

$$\omega = \frac{۱}{N}\sum_{i=۱}^{N} \mathrm{E}_{\mathcal{T}}\left[\hat{\psi}_i(y_i - \bar{y}_i)\right]$$

که اثبات کامل می‌شود. □

قضیه ۲.۱.۱. برای تابع زیان توان دوم خطا (تعریف ۱.۱.۱):

$$\omega = \frac{۲}{N}\sum_{i=۱}^{N} \mathrm{Cov}(\hat{Y}_i, Y_i)$$

که ω، خوش‌بینی موردانتظار (تعریف ۷.۱.۱)، N تعداد اعضای مجموعه مدل‌ساز $\mathcal{T}$، Y_i، i امین خروجی در مجموعه مدل‌ساز و $\hat{Y}_i = \hat{f}(\mathbf{X}_i)$، پیش‌بینی مدل از آن‌چه Y_i باید باشد، است.

برهان. فرض کنید در رابطه (۲.۱)،

$$\ell(x) = x(۱ - x)$$

باشد که تایید می‌کند تابع زیان توان دوم خطا متعلق به آن خانواده است.
بر اساس لم ۱.۱.۱، می‌توان نوشت

$$\omega = \frac{۱}{N}\sum_{i=۱}^{N} \mathrm{E}_{\mathcal{T}}\left[\hat{\psi}_i(Y_i - \bar{Y}_i)\right].$$

برای تابع زیان توان دوم خطا، همچنین

$$\hat{\psi}_i \equiv ۲\hat{Y}_i - ۱$$

و بنابراین

$$\begin{aligned}
\omega &= \frac{۱}{N}\sum_{i=۱}^{N} \mathrm{E}_{\mathcal{T}}\left[(۲\hat{Y}_i - ۱)(Y_i - \bar{Y}_i)\right] \\
&= \frac{۱}{N}\sum_{i=۱}^{N} \mathrm{E}_{\mathcal{T}}\left[(۲\hat{Y}_i - ۲\bar{\hat{Y}}_i + ۲\bar{\hat{Y}}_i - ۱)(Y_i - \bar{Y}_i)\right] \\
&= \frac{۱}{N}\sum_{i=۱}^{N} \mathrm{E}_{\mathcal{T}}\left[(۲\hat{Y}_i - ۲\bar{\hat{Y}}_i)(Y_i - \bar{Y}_i)\right] + \frac{۱}{N}\sum_{i=۱}^{N} \mathrm{E}_{\mathcal{T}}\left[(۲\bar{\hat{Y}}_i - ۱)(Y_i - \bar{Y}_i)\right]
\end{aligned}$$

که $\bar{\hat{Y}}_i = \mathrm{E}\left[\hat{f}(\mathbf{X}_i)\right]$ یک ثابت است (و بنابراین در محاسبات امید ریاضی در مرحله آخر حضور نداشت). سرانجام، دقت کنید که $\mathrm{E}_{\mathcal{T}}[Y_i] = \bar{Y}_i$ و بنابراین آخرین عبارت برابر صفر است. بنابراین

$$\begin{aligned}
\omega &= \frac{۱}{N}\sum_{i=۱}^{N} \mathrm{E}_{\mathcal{T}}\left[(۲\hat{Y}_i - ۲\bar{\hat{Y}}_i)(Y_i - \bar{Y}_i)\right] \\
&= \frac{۲}{N}\sum_{i=۱}^{N} \mathrm{Cov}(\hat{Y}_i, Y_i)
\end{aligned}$$

که اثبات کامل است. □

قضیه ۲.۱.۱ بیان می‌کند که خوش‌بینی خطای مدل‌ساز بیشتر از زمانی است که مدل یک کار خوب در پیش‌بینی متوسط y انجام می‌دهد که با نتیجه‌گیری «هر چه اریبی مدل کمتر باشد، واریانس مدل بیشتر و دقت خطای مدل‌ساز، کمتر می‌شود» کاملا سازگار است.

مطالب بیان شده حقیقت شگفت‌انگیزی را آشکار می‌سازند که حتی اگرچه کمینه کردن خطای مدل‌ساز منجر به مدلی با کمترین اریبی می‌شود، گاهی اوقات ممکن است برازش عمدی یک مدل اریب نسبت به دست‌یابی به مدلی با واریانس کمتر، مزیت‌هایی داشته باشد. علاوه بر این به‌نظر می‌رسد که هر روش ایجاد اریبی در مدل، پیچیدگی آن را کاهش می‌دهد، زیرا به‌نظر می‌رسد واریانس به پیچیدگی مدل ارتباط دارد.

گاهی‌اوقات در بعضی مدل‌های با بعد بسیار بالا، دلیل پیشین[۲۱] وجود دارد که برخی از

[21] Prior reason

متغیرها نباید در مدل باشند. برای مثال، آزمایش‌های میکروآرایه‌ای زیستی متعددی شامل هزاران ژن به‌عنوان پیشگوهای اولیه هستند که انتظار می‌رود تنها تعداد کمی از این ژن‌ها بر نتایج موثر باشند. در این موقعیت‌ها، انگیزه دیگری برای کاهش پیچیدگی مدل وجود دارد.

با این وجود، این سوال باقی می‌ماند که «چگونه باید پیچیدگی مدل را کاهش داد؟» یا به‌عبارت دیگر، «کدام متغیرها و چه تعدادی از آن‌ها باید از مدل حذف شوند؟». در ادامه، روش‌هایی برای پاسخ به این سوال مطرح می‌شود.

۱.۱.۱ مسیر پیش‌رو

مسیر پیش روی یک محقق، ساده و شفاف است ـ باید عبارتی پیدا شود که برای هر مدل دلخواه، به‌دقت Err را برآورد کند. سپس این اجازه داده می‌شود که مدلی انتخاب گردد که این عبارت را کمینه کند. به‌طور گسترده دو رهیافت وجود دارند که می‌توانیم اختیار کنیم

- روش‌هایی مثل اعتبارسنجی متقابل[۲۲] و تلاش‌های بوت‌استرپی[۲۳] در برآورد مستقیم Err بر اساس مجموعه مدل‌ساز $\mathcal{T}$.
- روش‌هایی مثل C_p مالو، معیار اطلاع آکائیک، معیار اطلاع بیزی و کمترین توان‌های دوم جریمه شده. در این روش‌ها، خوش‌بینی موردانتظار ω برآورد می‌شود و به $\overline{\text{err}}$ (که به‌سادگی بر اساس داده‌های درون مجموعه $\mathcal{T}$ قابل محاسبه است) اضافه می‌شود تا Err برآورد شود.

اعتبارسنجی متقابل

دلیل کم‌برآورد شدن Err توسط $\overline{\text{err}}$ این است که از داده‌های یکسانی به‌منظور هر دو هدف برازش مدل و نیکویی برازش استفاده شده است. اعتبارسنجی متقابل (استون، ۱۹۷۴)، (الن، ۱۹۷۱، که در آن از نام PRESS استفاده شده است) این مسئله را با تقسیم داده‌ها به K قسمت مجزا حل می‌کند. مدل را با استفاده از $(K-1)$تا از این قسمت‌ها برازش می‌دهد و نیکویی برازش را به کمک آخرین قسمت بیان می‌کند. به‌طور رسمی‌تر، امتیاز اعتبارسنجی متقابل[۲۴] به عنوان برآوردی برای خطای کلی موردانتظار تعریف می‌شود.

تعریف ۱.۱.۸. [امتیاز اعتبارسنجی متقابل] مجموعه مدل‌ساز $\mathcal{T}$ به K قسمت یا دسته مساوی تقسیم می‌شود. نماد $\hat{f}^{-\kappa(i)}$ برای نشان دادن مدل برازش شده به $\mathcal{T}$ بدون داده‌های قسمت i، نوشته می‌شود. بنابراین، امتیاز اعتبارسنجی‌ـ متقابل عبارت است از

$$\text{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\left(Y_i, \hat{f}^{-\kappa(i)}(X_i)\right)$$

[22] Cross Validation (CV)
[23] Bootstraping attempt
[24] Cross-validation score

آن‌چه باقی می‌ماند، چگونگی انتخاب K، تعداد دسته‌هایی است که باید داده‌ها به آن تقسیم شوند. دوباره موازنه اریبی–واریانس مطرح می‌شود. یک مقدار بزرگ K منجر به اریبی کم (به این دلیل که هر مجموعه تعداد زیادی از داده‌ها را شامل می‌شود) اما واریانس بزرگ (به دلیل اینکه مجموعه‌ها بسیار مشابه هستند) خواهد شد. در حالت $K = N$، مدل را N بار برازش می‌دهد، هر بار تنها یک داده حذف می‌شود ــ این حالت اعتبارسنجی متقابل تک حذفی[۲۵] نامیده می‌شود. در مقابل، مقدار کوچک K منجر به اریبی زیاد اما واریانس کم خواهد شد.

در راستای موازنه اریبی–واریانس،اغلب مقدار $K = ۵$ یا $K = ۱۰$ پیشنهاد شده است (کوهاوی، ۱۹۹۵) هم‌چنین، بریمان و اسپیکتور (۱۹۹۲)، نشان دادند که در بعضی موارد اعتبارسنجی متقابل ۵ تایی عملکرد بهتری از اعتبارسنجی متقابل تک حذفی دارد.

برای برازش مدل با استفاده از اعتبارسنجی متقابل، باید مدلی با کمترین امتیاز اعتبارسنجی متقابل انتخاب شود. هزینه محاسباتی اعتبارسنجی متقابل از اهمیت بسزایی برخوردار است به‌خصوص زمانی‌که K بزرگ است و مدل باید به تعداد دفعات زیادی برازش شود.

روش بوت‌استرپ

روش بوت‌استرپ توسط افرون (۱۹۷۹) تعریف شده است. ایده اصلی این روش، نمونه‌گیری باجای‌گذاری داده‌ها از مجموعه مدل‌ساز است که هر نمونه اندازه یکسانی مانند مجموعه مدل‌ساز اصلی داشته باشد (احتمالا شامل داده‌های تکراری نیز باشد). این کار، B بار انجام می‌شود و B مجموعه داده بوت‌استرپ تولید می‌شود.

چگونه می‌توان از مجموعه داده‌های بوت‌استرپ برای برآورد خطای پیش‌بینی استفاده کرد؟ یک رهیافت می‌تواند برازش مدل به هر مجموعه داده بوت‌استرپ باشد و چگونگی خوب پیش‌بینی کردن مجموعه مدل‌ساز اصلی بررسی شود. فرض کنید $\hat{f}^{*b}$ مدل به‌دست آمده از برازش داده‌ها در b امین مجموعه داده بوت‌استرپ باشد. بنابراین برآورد Err به‌صورت

$$\frac{۱}{B}\frac{۱}{N}\sum_{b=۱}^{B}\sum_{i=۱}^{N} L\left(Y_i, \hat{f}^{*b}(\mathbf{X}_i)\right)$$

محاسبه می‌شود. مسئله در این‌جا هم‌پوشانی بین مجموعه‌های بوت‌استرپ و نمونه‌هایی از مجموعه مدل‌ساز هستند که بر روی آن‌ها مدل آزمون شده است و این دقیقا همان دلیلی است که در اولین مطلب بیان شد یعنی $\overline{\text{err}}$ برآورد ضعیفی از Err را ارائه می‌داد.

مشابه روش اعتبارسنجی متقابل، برآورد بوت‌استرپ بهتری می‌تواند با آزمون مدل تنها بر روی داده‌هایی که در مجموعه متناظر بوت‌استرپ قرار ندارند، به‌دست آید.

تعریف ۹.۱.۱. [برآورد بوت‌استرپ Err] از مجموعه مدل‌ساز $\mathcal{T}$، B نمونه با جای‌گذاری تولید کنید. فرض کنید $\hat{f}^{*b}$ مدل برازش شده به b امین مجموعه داده بوت‌استرپ باشد. فرض کنید

[25] Leave-one out

C^{-i} مجموعه اندیس‌های b نمونه بوت‌استرپ باشد که مشاهده i را شامل نشود و $|C^{-i}|$ اندازه آن مجموعه باشد. برآورد بوت‌استرپ Err به‌صورت

$$\hat{\text{Err}}^{b} = \frac{۱}{N}\sum_{i=۱}^{N}\frac{۱}{|\dot{C}^{i}|}\sum_{b\in C^{-i}} L(Y_i, \hat{f}^{*b}(\mathbf{X}_i))$$

تعریف می‌شود.

این راه‌حل، پاسخی برای مسئله بیش‌برازشی ایجاد شده با روش اعتبارسنجی متقابل است. با این وجود، همچنان مسئله دیگری باقی مانده است و برای درک بیشتر آن، ابتدا قضیه ۳.۱.۱ اثبات می‌شود.

قضیه ۳.۱.۱ (افرون ، ۱۹۷۹)**.** احتمال اینکه یک مشاهده مشخص در هر نمونه بوت‌استرپ ظاهر شود، تقریبا ۰/۶۳۲ است. بنابراین، تعداد مشاهدات متمایز در هر نمونه بوت‌استرپ تقریبا N ۰/۶۳۲ است، که N، تعداد اعضای مجموعه مدل‌ساز است.

برهان. احتمال انتخاب هر عضو از مجموعه مدل‌ساز، $\frac{۱}{N}$ است. از این رو، احتمال انتخاب نشدن یک عضو، $(۱-\frac{۱}{N})$ است. از آن‌جایی که هر نمونه بوت‌استرپ، شامل N عضو است، برای اینکه یک نمونه خاص در نمونه بوت‌استرپ ظاهر نشود، باید N بار انتخاب نشود. بنابراین

$$P(i \text{ مشاهده} \notin b \text{ نمونه}) = \left(۱-\frac{۱}{N}\right)^{N}.$$

و از این رو

$$P(i \text{ مشاهده} \in b \text{نمونه}) = ۱-\left(۱-\frac{۱}{N}\right)^{N}.$$

برای N بزرگ

$$P(i \text{ مشاهده} \in b \text{ نمونه}) \approx ۱-\exp\{-۱\} = ۰/۶۳۲.$$

بدین ترتیب، میانگین تعداد مشاهدات متمایز در هر نمونه بوت‌استرپ تقریبا N ۰/۶۳۲ است. □

قضیه ۳.۱.۱ نتیجه آن است که برآورد بوت‌استرپ Err تقریبا شبیه اعتبارسنجی متقابل دوتایی ($K=۲$) رفتار می‌کند، زیرا هر نمونه بوت‌استرپ تقریبا شامل نیمی از داده‌ها می‌شود. با این حال، همان‌طور که اشاره شد، $K=۵$ یا $K=۱۰$ بهینه است. بنابراین، نمونه‌های بوت‌استرپ کوچک‌تر از آن چه که ممکن است ایده‌آل باشد، هستند. از این رو، مدل‌های به‌دست آمده پیچیدگی کمتری دارند و خطای پیش‌بینی، اریبی رو به افزایش خواهد داشت.

برای رفع این مشکل، افرون (۱۹۸۳) برآوردگر

$$\hat{\text{Err}}^{(۰/۶۳۲)} = ۰/۳۶۸\cdot\overline{\text{err}} + ۰/۶۳۲\cdot\hat{\text{Err}}^{b} \qquad (۴.۱)$$

را پیشنهاد داد. این مدل، ترکیبی از مشاهداتی که خیلی نزدیک به مجموعه مدل‌ساز هستند (برای محاسبه $\overline{\text{err}}$ به‌کار گرفته شده‌اند) و نقاطی که بسیار دور از مجموعه مدل‌ساز هستند (برای محاسبه $\hat{\text{Err}}^b$ استفاده شده‌اند) است تا میانگین متعادلی را نتیجه دهد.

بحث مطرح شده تاکنون بسیار کلی بوده است. از مدل‌های متعددی که در یادگیری آماری استفاده می‌شوند اساسی‌ترین و پرکاربردترین آن‌ها، مدل خطی است.

۲.۱ مدل خطی

در مدل خطی، توابعی به‌صورت

$$\hat{f}(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \tag{۵.۱}$$

مورد مطالعه هستند. در کلی‌ترین شکل آن، مدل خطی صورت $\hat{f}(\mathbf{X}) = \beta_0 + \mathbf{X}\boldsymbol{\beta}$ را به‌خود می‌گیرد (که β_0 ممکن است صفر باشد). با این حال، داده‌ها می‌توانند نرمال‌سازی شوند تا به عبارت ثابت β_0 نیاز نباشد که به‌طور قابل توجهی نمادگذاری را ساده می‌کند. هم‌چنین نرمال‌سازی بردارهای $\mathbf{X}$ به‌منظور داشتن متغیرهایی با میانگین صفر و واریانس یک، مفید است. تعریف ۱.۲.۱ شرایط نرمال‌سازی را بیان می‌کند.

تعریف ۱.۲.۱ (شرایط نرمال‌سازی). همیشه فرض می‌شود که ورودی‌های $\mathbf{X}$ و خروجی‌های $\mathbf{Y}$ به‌صورت زیرنرمال‌سازی شده‌اند.

$$\sum_{i=1}^{N} Y_i = 0 \qquad \sum_{i=1}^{N} X_{ij} = 0, \forall j \qquad \sum_{i=1}^{N} X_{ij}^2 = 1, \forall j$$

در مدل‌های خطی،

- یک مدل ساده، مدلی است که از هیچ یک از متغیرهای ماتریس $\mathbf{X}$ برای برآورد $\mathbf{Y}$ استفاده نمی‌کند. بنابراین در یک مدل ساده، تعداد مولفه‌های بردار $\boldsymbol{\beta}$ صفر خواهند بود.
- در یک مدل پیچیده، بیشتر مولفه‌های بردار پارامتر $\boldsymbol{\beta}$ غیرصفر خواهند بود که این حقیقت را بیان می‌کند که از بیشتر متغیرها استفاده شده است.

هدف یک آماردان، یافتن بردار $\boldsymbol{\beta}$ ای است که خطای کلی موردانتظار (تعریف ۲.۱.۱) را کمینه می‌کند.

ابتدا رهیافت ساده کمینه‌سازی خطای مدل‌ساز (تعریف ۴.۱.۱) مطرح می‌شود. این روش، کمترین توان‌های دوم معمولی[۲۶] (OLS) نامیده می‌شود. در مدل خطی، می توان نوشت:

$$\overline{\text{err}}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{f}(\mathbf{X}_i)\right)^2$$

[26] Ordinary Least Squares (OLS)

$$\begin{aligned} &= \frac{۱}{N}\sum_{i=۱}^{N}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^۲ \\ &= \frac{۱}{N}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^۲ \end{aligned} \tag{۶.۱}$$

که در آن $||.||$ نرم با مشتق گرفتن از رابطه (۶.۱) و برابر قرار دادن آن با صفر، می‌توان نوشت

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-۱}\mathbf{X}^\top\mathbf{Y}. \tag{۷.۱}$$

به ازای بردار ورودی $\mathbf{x}$ مشخص، برآورد متناظر y عبارت است از

$$\hat{y}^{\text{OLS}} = \hat{f}(\mathbf{x}) = \mathbf{x}^\top\hat{\boldsymbol{\beta}}^{\text{OLS}} = \mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-۱}\mathbf{X}^\top y \tag{۸.۱}$$

ماتریس هَت[۲۷] به‌صورت زیر تعریف می‌شود.

تعریف ۲.۲.۱ (ماتریس هَت).

$$\boldsymbol{H}^{\text{OLS}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-۱}\mathbf{X}^\top.$$

این رابطه، اجازه می‌دهد خروجی‌های مجموعه مدل‌ساز $\mathbf{Y}$ و خروجی‌های مدل $\hat{f}(\mathbf{X})$ به‌صورت

$$\hat{\mathbf{Y}}^{\text{OLS}} = \boldsymbol{H}^{\text{OLS}}\mathbf{Y}$$

با یکدیگر ارتباط داشته باشند.

۱.۲.۱ توزیع $\hat{\boldsymbol{\beta}}$

اگر خطاهای ϵ از توزیع نرمال پیروی کنند، آن‌گاه

$$\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma_\epsilon^۲\boldsymbol{I})$$

که در آن $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$، نرمال چند متغیره d بعدی با بردار میانگین $\boldsymbol{\mu}$ و ماتریس کوواریانس $\boldsymbol{\Sigma}$ است.

همچنین $\hat{\boldsymbol{\beta}}^{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-۱}\mathbf{X}^\top\mathbf{Y}$. یک تبدیل خطی از $\mathbf{Y}$ است. زیرا در توزیع نرمال چندمتغیره اگر $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ و $\mathbf{X} = \boldsymbol{A}\mathbf{Y}$، آن‌گاه $\mathbf{X} \sim N_n(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$. بنابراین،

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\text{OLS}} &\sim N_p\left((\mathbf{X}^\top\mathbf{X})^{-۱}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}, (\mathbf{X}^\top\mathbf{X})^{-۱}\mathbf{X}^\top\boldsymbol{I}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-۱}\sigma_\epsilon^۲\right) \\ &\sim N_p\left(\boldsymbol{\beta}, (\mathbf{X}^\top\mathbf{X})^{-۱}\sigma_\epsilon^۲\right). \end{aligned} \tag{۹.۱}$$

توجه کنید که

[27] Hat matrix

- با فرض برقراری شرایط نرمال‌سازی در تعریف ۱.۲.۱ ، ماتریس واریانس‌ـ کوواریانس داده‌های ورودی، $\mathbf{X}$، به‌صورت

$$\boldsymbol{S} = \frac{\mathbf{X}^\top \mathbf{X}}{N}$$

است. از این رو، ماتریس واریانس‌ـ کوواریانس برآوردگر $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ متناسب با معکوس $\boldsymbol{S}$ است.

واریانس بالا، بدین معناست که داده‌ها در مجموعه $\mathcal{T}$ در طول آن جهت خاص بسیار پراکنده شده‌اند. در نتیجه، ماهیت مدل در راستای آن جهت به‌خوبی تعریف شده است و بنابراین واریانس برآورد شیب آن خط بسیار کم خواهد بود. (شکل ۲.۱ را برای نمایش این پدیده ببینید).

- $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ و بنابراین، $\mathrm{E}(\hat{Y}) = Y$. به‌عبارت دیگر، مدل خطی، نااریب است.

در حقیقت، بر اساس قضیه گوس‌ـ مارکوف[۲۸] (کاریا و کوراتا ، ۲۰۰۴)، برآوردگر کمترین توان‌های دوم معمولی کمترین واریانس را در بین همه برآوردگرهای نااریب خطی دارند، دارد. با این وجود، همچنان که قبلا اشاره شد، واریانس می‌تواند با افزایش مقداری اریبی، کاهش پیدا کند.

۲.۲.۱ تفسیر هندسی کمترین توان‌های دوم معمولی

برآورد کمترین توان‌های دوم $\mathbf{Y}$ را درنظر بگیرید:

$$\hat{\mathbf{Y}}^{\mathrm{OLS}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}.$$

به‌وضوح، طرف سمت راست این معادله یک ترکیب خطی از ستون‌های $\mathbf{X}$ است. به‌عبارت دیگر، برآورد $\hat{\mathbf{Y}}$ در فضای برداری تولید شده با ستون‌های $\mathbf{X}$ قرار می‌گیرد. پس، کمترین توان‌های دوم معمولی به‌سادگی، یافتن نزدیک‌ترین بردار به $\mathbf{Y}$ است که در آن فضای برداری قرار می‌گیرد که می‌توان به‌عنوان تعمیمی از تجزیه مقدار ویژه [۲۹] برای ماتریس‌های غیرمربعی تصور کرد.

ماتریس $\mathbf{X}$ یک ماتریس $N \times p$ است که می‌تواند به‌عنوان تبدیل یک بردار از فضای $\boldsymbol{\beta}$ (p بُعدی) به فضای $\hat{\mathbf{Y}}$ (N بُعدی) درنظر گرفته شود، یک $\boldsymbol{\beta}$ را می‌گیرد و آن را به $\hat{\mathbf{Y}}$ درست تبدیل می‌کند. بر اساس تعریف مطرح شده در هیستی و همکاران (۲۰۰۹)، تجزیه مقدار ویژه ماتریس $\mathbf{X}$ عبارت است از

$$\mathbf{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$$

که

[28] Gauss-Markov theorem

[29] Eigenvalue decomposition

- $\boldsymbol{V}$ یک ماتریس $p \times p$ است که ستون‌هایش متعامد و فضای سطری $\mathbf{X}$ را تولید می‌کند.

  همچنین می‌تواند به‌عنوان تبدیلی که هرآنچه بردار $\mathbf{X}$ از فضای $\boldsymbol{\beta}$ به درون فضای ویژه $\mathbf{X}$ تبدیل می‌شود، درنظر گرفته شود.

  توجه کنید که $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}$.

- $\boldsymbol{D}$ یک ماتریس قطری $p \times p$ شامل مولفه‌های $d_1, \ldots, d_p$ است که ماتریس مقادیر تکین[۳۰] $\mathbf{X}$ نامیده می‌شود که مشابه مقادیر ویژه $\mathbf{X}$ هستند.

- $\boldsymbol{U}$ یک ماتریس $N \times p$ است که ستون‌هایش متعامد و فضای ستونی $\mathbf{X}$ را تولید می‌کند و می‌تواند به‌عنوان تبدیلی که برداری را از فضای ویژه $\mathbf{X}$ به فضای $\hat{\mathbf{Y}}$ برمی‌گرداند، درنظر گرفته شود.

  توجه کنید که $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$، اما از آن‌جایی‌که ماتریس مربعی نیست، $\boldsymbol{U}\boldsymbol{U}^\top \neq \boldsymbol{I}$.

با توجه به اینکه

$$\hat{\mathbf{Y}}^{\text{OLS}} = \boldsymbol{H}^{\text{OLS}}\mathbf{Y}$$

طبق آن‌چه مطرح شد، ماتریس $\boldsymbol{H}$ باید $\mathbf{Y}$ را به درون فضای تولید شده با ستون‌های $\mathbf{X}$، تصویر کند. برای درک بیشتر، صورت مقدار_تکین ماتریس $\boldsymbol{H}$ را درنظر بگیرید.

$$\begin{aligned}
\boldsymbol{H}^{\text{OLS}} &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \\
&= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \\
&= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{D}\boldsymbol{D}\boldsymbol{V}^\top)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \\
&= \boldsymbol{U}\boldsymbol{U}^\top \qquad (۱۰.۱)
\end{aligned}$$

در این حالت، به‌راحتی ملاحظه می‌شود که $\hat{\boldsymbol{H}}^{\text{OLS}}$ در واقع، یک ماتریس تصویر است. ماتریس $\boldsymbol{U}^\top$ ابتدا $\mathbf{Y}$ را درون فضای ویژه $\mathbf{X}$ تصویر می‌کند (توجه کنید که $\boldsymbol{U}^\top\mathbf{Y}$ مختصات‌های $\mathbf{Y}$ بیان شده بر اساس پایه $\boldsymbol{U}$ است) و سپس دوباره بردار نتیجه شده را به درون فضای $\mathbf{Y}$ (بااستفاده از $\boldsymbol{U}$) برمی‌گرداند. نتیجه، یک تقریب از $\mathbf{Y}$ تنها با استفاده از ستون‌های $\boldsymbol{U}$ است.

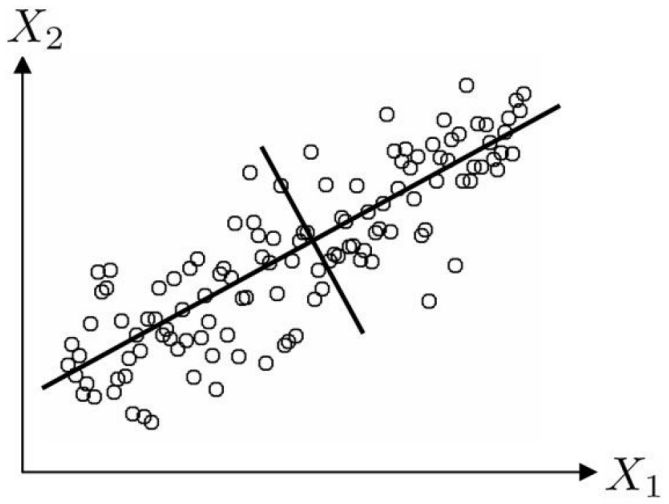
در راستای درک بهتری از مقادیر تکین، ماتریس واریانس_کوواریانس $\mathbf{X}$ به شکل مقدار_تکین به‌صورت

$$\begin{aligned}
\boldsymbol{S} &= \frac{\mathbf{X}^\top\mathbf{X}}{N} \\
&= \frac{\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top}{N} \\
&= \frac{\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top}{N}
\end{aligned}$$

[30] Singular values

است که همان تجزیه‌ـویژه ماتریس $\mathbf{X}^\top\mathbf{X}$ می‌باشد. بنابراین، مقادیر تکین d_j (جذر) مقادیر ویژه $\boldsymbol{S}$ هستند و بردارهای $\mathbf{v}_j$ که ستون‌های ماتریس $\boldsymbol{V}$ را می‌سازند، بردارهای ویژه $\boldsymbol{S}$ هستند.

فرض کنید d_1 بزرگترین مقدار تکین باشد. این بدین معناست که $\mathbf{v}_1$ ترکیب خطی ستون‌های $\mathbf{X}$ است که واریانس نمونه‌ای بزرگتری از سایر بردارهای ویژه دارد. آن‌گاه، بنابر ویژگی عمود بودن بر $\mathbf{v}_1$، $\mathbf{v}_2$ دومین مقدار تکین بزرگ است (ماتریس $\boldsymbol{V}$، یک ماتریس متعامد است) و به همین صورت ادامه می‌یابد. این جهت‌ها، مولفه‌های اصلی [۳۱] ماتریس $\mathbf{X}$ نامیده می‌شوند. کمترین توان‌های دوم معمولی، بردار $\mathbf{Y}$ را بر روی این مولفه‌ها تصویر می‌کند.این مفهوم در شکل ۲.۱ نشان داده شده است.



شکل ۲.۱: مولفه‌های اصلی ماتریس $\mathbf{X}$ شامل دو ستون مولفه‌های اصلی ماتریس $\mathbf{X}$ شامل دو ستون ـ یکی برای X_1 و یکی برای X_2. محور Y در بیرون از صفحه قرار دارد و نقطه‌ها تصویرهای نقاط داده‌ها روی صفحه $X_1 - X_2$ هستند. خط‌های پررنگ نشان‌دهنده مولفه‌های اصلی ماتریس هستند ـ اولین مولفه اصلی دارای بزرگترین مقدار تکین است و دومی، بر اولی عمود است. کمترین توان‌های دوم معمولی، بردار $\mathbf{Y}$ را بر این مولفه‌ها تصویر می‌کند.

۳.۲.۱ خوش‌بینی موردانتظار در عمل

در این بخش، حالت خوش‌بینی در حالت خاص مدل‌های خطی ارا رائه می‌کنیم:

قضیه ۱.۲.۱. برای یک روش پیش‌بینی که در رابطه $\hat{\mathbf{Y}} = \boldsymbol{H}\mathbf{Y}$ صدق می‌کند، خوش‌بینی موردانتظار به‌صورت

$$\omega = \frac{2}{N}\operatorname{tr}(\boldsymbol{H})\sigma_\epsilon^2$$

است.

[31] Principle components

برهان. در مدل‌های خطی، $\hat{\mathbf{Y}} = \boldsymbol{H}\mathbf{Y}$ که $\boldsymbol{H}$ ماتریس هت (تعریف ۲.۲.۱) است. هم‌چنین، فرض کنید $\boldsymbol{H}_i$، iامین سطر از ماتریس $\boldsymbol{H}$ را نشان دهد. آن‌گاه

$$\hat{Y}_i = \boldsymbol{H}_i\mathbf{Y} = \sum_{k=1}^{N} H_{ik}Y_k$$

بنابراین،

$$\begin{aligned}\omega &= \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\hat{Y}_i, Y_i)\\ &= \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\sum_{k=1}^{N}H_{ik}Y_k, Y_i)\end{aligned}$$

از آن‌جایی که Y_iها مستقل هستند، فقط یک کوواریانس باقی می‌ماند، به‌عبارت دیگر،

$$\begin{aligned}\omega &= \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\boldsymbol{H}_{ii}Y_i, Y_i)\\ &= \frac{2}{N}\sum_{i=1}^{N}H_{ii}\mathrm{V}(Y_i)\\ &= \frac{2}{N}\mathrm{tr}(\boldsymbol{H})\sigma_\epsilon^2 \qquad (11.1)\end{aligned}$$

و اثبات کامل می‌شود. □

قضیه ۱.۲.۱ بیان می‌کند که اثر ماتریس $\boldsymbol{H}$ مستقیما با پیچیدگی مدل ارتباط دارد. ماتریس $\boldsymbol{H}$، یک تصویر از $\mathbf{Y}$ به فضای کوچک‌تری است. بیان $\mathbf{Y}$ به صورت قطری منجر به یک ماتریس می‌شود که اثر آن با بعد فضایی است که ماتریس $\boldsymbol{H}$ تصویر می‌کند. این موضوع، انگیزه تعریف تعداد پارامترهای مؤثر[۳۲] است.

تعریف ۳.۲.۱. (تعداد پارامترهای مؤثر)

$$d_{\text{eff}} = \mathrm{tr}(\boldsymbol{H}).$$

قضیه ۲.۲.۱. اگر یک مدل خطی با استفاده از برآوردگر کمترین توان‌های دوم (به‌عبارت دیگر، $\hat{f}(\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{OLS}}$) برازش شود، آن‌گاه

$$\omega = \frac{2p}{N}\sigma_\epsilon^2 \qquad (12.1)$$

که p تعداد متغیرها در $\mathbf{X}$ و $\sigma_\epsilon^2 = \mathrm{V}(Y_i)$، خطای ساده‌نشدنی در مدل موردنظر است.

[32] Effective number of parameters

برهان. ابتدا ثابت می‌شود که در این حالت خاص، $\operatorname{tr}(\boldsymbol{H}) = p$. از آن‌جایی که $\boldsymbol{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ و $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{AB})$، بنابراین می‌توان نوشت:

$$\begin{aligned}\operatorname{tr}(\boldsymbol{H}) &= \operatorname{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\\ &= \operatorname{tr}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1})\\ &= \operatorname{tr}(\ I) = P. \qquad (۱۳.۱)\end{aligned}$$

از ترکیب رابطه‌های (۱۱.۱) و (۱۳.۱)، می‌توان نتیجه گرفت

$$\omega = \frac{2p}{N}\sigma_\epsilon^2$$

و اثبات کامل می‌شود. □

۴.۲.۱ طرح متعامد

در ادامه، حالتی در نظر گرفته می‌شود که $\mathbf{X}^\top\mathbf{X} = \boldsymbol{I}$ که آن را طرح متعامد[۳۳] می‌نامند و منجر به ساده‌تر شدن رابطه (۶.۱) می‌شود.

قضیه ۳.۲.۱. در حالت طرح متعامد که $\mathbf{X}^\top\mathbf{X} = \boldsymbol{I}$، رابطه (۶.۱) می‌تواند به‌صورت

$$Q(\boldsymbol{\beta}) = \frac{1}{N}\left\|\mathbf{Y} - \hat{\mathbf{Y}}^{\mathrm{OLS}}\right\|^2 + \frac{1}{N}\left\|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}\right\|^2 \qquad (۱۴.۱)$$

نوشته شود که در آن $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ برآورد کمترین توان‌های دوم معمولی برای $\boldsymbol{\beta}$ و $\hat{\mathbf{Y}}^{\mathrm{OLS}}$ برآورد کمترین توان‌های دوم معمولی برای $\mathbf{Y}$ است.

برهان.

$$\begin{aligned}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 &= \mathbf{Y}^\top\mathbf{Y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\\ &= \mathbf{Y}^\top\mathbf{Y} - 2\mathbf{Y}^\top\mathbf{X}\mathbf{X}^\top\mathbf{Y} + \mathbf{Y}^\top\mathbf{X}\mathbf{X}^\top\mathbf{Y} + \mathbf{Y}^\top\mathbf{X}\mathbf{X}^\top\mathbf{Y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y}\\ &\quad +\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\\ &= \mathbf{Y}^\top\mathbf{Y} - 2\mathbf{Y}^\top\mathbf{X}\mathbf{X}^\top\mathbf{Y} + \mathbf{Y}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{Y} + \mathbf{Y}^\top\mathbf{X}\mathbf{X}^\top\mathbf{Y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y}\\ &\quad +\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\\ &= \mathbf{Y}^\top(\boldsymbol{I} - \mathbf{X}\mathbf{X}^\top)^\top(\boldsymbol{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{Y} + (\mathbf{X}^\top\mathbf{Y} - \boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{Y} - \boldsymbol{\beta})\\ &= \left\|\mathbf{Y} - \hat{\mathbf{Y}}^{\mathrm{OLS}}\right\| + \left\|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}\right\|\end{aligned}$$

و اثبات کامل است. □

این قضیه تقریبا بدیهی به‌نظر می‌رسد ــ و در واقع، در حالت کمترین توان‌های دوم معمولی برقرار است. اهمیت قضیه ۳.۲.۱ این است که اولین عبارت در رابطه (۱۴.۱) به $\boldsymbol{\beta}$، آرگومان

[33] Orthonormal design

تابع Q، بستگی ندارد. این تابع به $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ بستگی دارد که فقط به داده‌های ورودی بستگی دارد. این بدین معنی است که کمینه کردن $Q(\boldsymbol{\beta})$ نسبت به $\boldsymbol{\beta}$، فقط به عبارت دوم نیاز است که می‌تواند به‌صورت $\sum_{j=۱}^{p}(\hat{\beta}_j^{\text{OLS}} - \beta_j)^۲$ نوشته شود. در این فرم، بدیهی است که کافی است هر مولفه $\boldsymbol{\beta}$ به‌طور جداگانه کمینه شود.

به‌عبارت دیگر، کمینه کردن $Q(\boldsymbol{\beta})$ نسبت به $\boldsymbol{\beta}$ هم‌ارز با کمینه کردن

$$q(\beta_j) = \left(\hat{\beta}_j^{\text{OLS}} - \beta_j\right)^۲$$

برای هر β_j است.

در این حالت خاص، بدیهی است که پاسخ $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{\text{OLS}}$ است. در برخی مدل‌های پیچیده‌تر، چنین نخواهد بود.

۵.۲.۱ اعتبارسنجی متقابل در مدل خطی

محاسبات اعتبارسنجی متقابل، در بعضی مسائل خاص می‌تواند به‌سرعت انجام شود. برای مثال، هنگامی‌که یک مدل خطی به صورت $\hat{\mathbf{Y}} = \boldsymbol{H}\mathbf{Y}$ برآورد می‌شود، هیستی و تیبشیرانی (۱۹۹۰) نشان دارند که امتیاز اعتبارسنجی متقابل به‌صورت

$$\text{CV}\left(\hat{f}\right) = \frac{۱}{N}\sum_{i=۱}^{N}\left(\frac{Y_i - \hat{f}(X_i)}{۱ - H_{ii}}\right)^۲$$

محاسبه می شود. در بعضی موارد، محاسبه اثر ماتریس $\boldsymbol{H}$ از محاسبه عنصر قطری iام آن آسان‌تر است و این منجر به برآوردی از امتیاز اعتبارسنجی متقابل می‌شود که امتیاز اعتبارسنجی متقابل تعمیم‌یافته[۳۴] نامیده می‌شود و به‌صورت

$$\text{GCV}\left(\hat{f}\right) = \frac{۱}{N}\sum_{i=۱}^{N}\left(\frac{y_i - \hat{f}(X_i)}{۱ - \text{tr}(\boldsymbol{H})/N}\right)^۲ \qquad (۱۵.۱)$$

برآورد Err با خوش‌بینی موردانتظار

در ادامه روش‌هایی در نظر گرفته شده که در آن‌ها Err با خوش‌بینی موردانتظار ω برآورد می‌شود.

C_p مالو و AIC

در قضیه ۲.۲.۱، مطرح شد که برای برآوردگر کمترین توان‌های دوم معمولی، خوش‌بینی موردانتظار به‌صورت

$$\omega = \frac{۲p}{N}\sigma_\epsilon^۲$$

است. آماره C_p (مالو، ۱۹۷۳) به‌سادگی این ω را به خطای مدل‌ساز اضافه می‌کند تا یک برآورد برای خطای کلی موردانتظار به‌دست آورد.

[34] Generalized Cross-Validation (GCV)

تعریف ۴.۲.۱ (C_p مالوز).

$$C_p = \frac{۱}{N}\left\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|^۲ + ۲\frac{p}{N}\sigma_\epsilon^۲$$

که در آن، N تعداد مشاهدات در مجموعه مدل‌ساز و p تعداد متغیرها در مدل است.

واریانس $\sigma_\epsilon^۲$ با استفاده از یک روش اریبی کم (مانند کمترین توان‌های دوم معمولی) برای برازش داده‌ها برآورد می‌شود. سپس، برای هر مدل پیشنهادی، C_p محاسبه می‌شود و مدل با کمترین مقدار انتخاب می‌شود.

معیار اطلاع آکائیک (AIC) (آکائیک، ۱۹۷۴) به روش مشابهی عمل می‌کند اما از تابع زیان تاحدی متفاوت استفاده می‌کند که آن را برای خانواده کلی‌تری از مدل‌ها، کاربردی‌تر می‌سازد. در حالت مدل خطی با خطاهای نرمال، معیارهای AIC و C_p هم‌ارز هستند.

در ادامه، یک نتیجه مفید اثبات می‌شود.

نتیجه ۱.۲.۱. در مدل‌های خطی کمینه کردن معیار AIC و کمینه‌کردن امتیاز اعتبارسنجی متقابل تعمیم‌یافته (GCV)، به‌طور مجانبی معادل هستند.

برهان. برای اثبات، امتیاز GCV در رابطه (۱۵.۱) در نظر بگیرید. با استفاده از تقریب $\frac{۱}{(۱-x)^۲} \approx ۱ + ۲x$،

$$\mathrm{GCV}\left(\hat{f}\right) \approx \frac{۱}{N}\sum_{i=۱}^{N}\left(y_i - \hat{f}(x_i)\right)^۲\left(۱ + ۲\frac{\mathrm{tr}(\boldsymbol{H})}{N}\right).$$

در مدل خطی، ملاحظه شد که $\mathrm{tr}(\boldsymbol{H}) = p$ و بنابراین،

$$\mathrm{GCV}\left(\hat{f}\right) = \frac{۱}{N}||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^۲ + \frac{۲P}{N}\left\{\frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^۲}{N}\right\}. \qquad (۱۶.۱)$$

عبارت داخل آکولاد به‌سادگی یک برآورد از واریانس $\sigma_\epsilon^۲$ است و بنابراین، رابطه (۱۶.۱) همان AIC است! □

نتیجه ۱.۲.۱ تایید می‌کند که روش‌های متفاوتی برای انجام یک کار، یعنی برآورد کردن خطای کلی مورد انتظار فقط با استفاده از داده‌ها در مجموعه مدل‌ساز، وجود دارند. این نتیجه برای اعتبارسنجی متقابل نیز (به‌جای اعتبارسنجی متقابل تعمیم‌یافته) می‌تواند به کار رود (استون، ۱۹۷۷).

BIC

معیار اطلاع بیزی (BIC) (شوارتز، ۱۹۷۸) بسیار شبیه به معیارهای C_p و AIC است.

تعریف ۵.۲.۱ (معیار اطلاع بیزی). برای مدل خطی با خطاهای نرمال،

$$\mathrm{BIC} = \frac{N}{\sigma_\epsilon^۲}\left[\frac{۱}{N}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^۲ + (\log N)\frac{p}{N}\sigma_\epsilon^۲\right]$$

که در آن، N تعداد مشاهدات در مجموعه مدل‌ساز و p تعداد متغیرها در مدل است.

توجه کنید که تنها تفاوت بین تعریف‌های ۴.۲.۱ و ۵.۲.۱ عبارت $\log N$ به‌جای عدد ۲ است. برای $N > e^{۲} \approx ۷/۴$، $\log N > ۲$ و BIC پیچیدگی مدل را با شدت بیشتری از AIC جریمه می‌کند.

همانند AIC، معیار BIC برای هر مدل پیشنهادی محاسبه می‌شود و مدل با کمترین BIC انتخاب می‌شود. برخلاف شباهت با AIC، معیار BIC از اصول کاملا متفاوت، روش‌های بیزی، به‌دست آمده است. در ادامه، به یک روش ابتکاری کوتاه برای چگونگی به‌دست آوردن معیار BIC پرداخته شده است. برای جزئیات بیشتر، شوارتز (۱۹۷۸) را ببینید. (هم‌چنین، کاس و رافتری، ۱۹۹۵ چگونگی به‌دست آوردن این معیار را ارائه کرده‌اند. کاواناگ، ۲۰۰۹ نیز روش دست‌یابی به این معیار را با جزئیات کمتری بیان کرده است).

فرض کنید یک مجموعه از مدل‌های پیشنهادی $\mathcal{M} = \{\mathcal{M}_۱, \cdots, \mathcal{M}_m\}$ وجود داشته باشد که هر کدام متناظر با مدل‌هایی با پارامترهای θ_m با تعداد p_m پارامتر است. هدف، انتخاب بهترین مدل در بین آن‌ها بر اساس داده‌ها در مجموعه مدل‌ساز $\mathcal{T}$ است. رهیافت بیزی برای این مسئله، یافتن مدل $\mathcal{M}$ است که

$$P(\mathcal{M}_n|\mathcal{T})$$

را ماکسیمم کند. با استفاده از قضیه بیز، آن می‌تواند به‌صورت

$$\begin{aligned} P(\mathcal{M}_n|\mathcal{T}) &= P(\mathcal{M}_n) \cdot P(\mathcal{T}|\mathcal{M}_n) \\ &= P(\mathcal{M}_n) \int P(\mathcal{T}|\theta_n, \mathcal{M}_n) P(\theta_n|\mathcal{M}_n) \mathrm{d}\theta_n. \end{aligned} \tag{۱۷.۱}$$

در اینجا، $P(\mathcal{M}_n)$ یک توزیع پیشین روی همه مدل‌های ممکن $\mathcal{M}_m$ است، اکنون انتگرال را در نظر بگیرید.

- اولین عبارت در انتگرال، درستنمایی، احتمال به‌دست آوردن داده‌ها در مجموعه مدل‌ساز به شرط یک مدل مشخص و پارامتر مرتبط، است.
- عبارت دوم، احتمال به‌دست آوردن یک پارامتر خاص به شرط یک مدل معین است. به‌نوعی، این یک نوع احتمال پیشین برای پارامترهای یک مدل معین است.

بنا به تعریف برآوردگر ماکسیمم درستنمایی[۳۵] (MLE)، $\hat{\theta}_n$ یک MLE برای پارامتر θ است که درستنمایی $P(\mathcal{T}|\theta_n, \mathcal{M}_n)$ را ماکسیمم می‌کند. اکنون، اگر فرض شود که تنها پارامتری که برای هر مدل معین باقی می‌ماند با $\hat{\theta}_n$ جایگزین می‌شود، آن‌گاه $P(\theta_n|\mathcal{M}_n) = \mathbb{1}_{\theta_n=\hat{\theta_n}}$ و انتگرال بالا به‌سادگی برابر با درستنمایی در نقطه $\hat{\theta}_n$، یعنی $P(\tau|\hat{\theta}_n, \mathcal{M}_n)$ می‌شود.

برآوردگر ماکسیمم درستنمایی واریانس غیر صفر دارد و بنابراین نمی‌توان با اطمینان کامل بیان کرد که $\theta = \hat{\theta}$. دیگر پارامترهای نزدیک نیز قابلیت نگهداری در مدل را دارند که منجر به کوچکتر کردن انتگرال می‌شود (زیرا وزنی به درستنمایی‌های کوچکتر داده شده است).

[35] Maximum Likelihood Estimator (MLE)

این موضوع با استفاده از بسط لاپلاس می‌تواند فرمول‌بندی شود. اما، ابتدا تعاریفی برای درک بهتر آن ارائه می‌شود.

تعریف ۶.۲.۱ (limsup)**.** دنباله‌ی $\{X_n\}$ را در نظر بگیرید. در این صورت

$$\begin{aligned}\limsup_{n\to\infty} X_n &= \lim_{n\to\infty}\left(\sup_{m\geq n} X_m\right)\\ &= \inf_{n\geq ۰}\sup_{m\geq n} X_m\\ &= \inf\left\{\sup\left\{X_m : m\geq n\right\} : n\geq ۰\right\}.\end{aligned}$$

یا به‌طور معادل، می‌توان فرض کرد دنباله‌ی J_n به‌صورت زیر تعریف شده باشد:

$$\begin{aligned}J_n &= \sup\left\{X_m : m\in\{n, n+۱, n+۲, \ldots\}\right\}\\ &= \cup_{m=n}^{\infty} X_m = X_n \cup X_{n+۱}\cup X_{n+۲}\cup\cdots\end{aligned}$$

دنباله‌ی J_n یک دنباله‌ی ناصعودی ($J_n \supseteq J_{n+۱}$) است، زیرا $J_{n+۱}$ اجتماع مجموعه‌های کوچکتر از J_n است. بزرگترین کران پایین روی این دنباله از اجتماع X_nها برابر است با

$$\begin{aligned}\limsup_{n\to\infty} X_n &= \inf\left\{\sup\left\{X_m : m\in\{n, n+۱, n+۲, \ldots\}\right\} : n\in\{۱, ۲, \ldots\}\right\}\\ &= \cap_{n=۱}^{\infty}\left(\cup_{m=n}^{\infty} X_m\right)\end{aligned}$$

تعریف ۷.۲.۱ (نماد O بزرگ[۳۶])**.** فرض کنید X_n و Y_n دو دنباله از متغیرهای تصادفی باشند. اگر

$$\limsup_{n\longrightarrow\infty} |\frac{X_n}{Y_n}| < \infty$$

یا به عبارت دیگر اگر $n\longrightarrow\infty$، عبارت $|\frac{X_n}{Y_n}|$ کران دار باشد، می‌نویسیم $X_n = O_p(Y_n)$

تعریف ۸.۲.۱ (نماد o کوچک[۳۷])**.** فرض کنید X_n و Y_n دو دنباله از متغیرهای تصادفی باشند. اگر

$$\lim_{n\longrightarrow\infty}\left|\frac{X_n}{Y_n}\right| = ۰$$

می‌نویسیم $X_n = o_p(Y_n)$

در ریاضیات علامت O بزرگ رفتار حدی یک تابع، وقتی آرگومان های آن به یک عدد خاص میل می‌کند را بیان می‌کند، علامت O بزرگ به کاربر اجازه می‌دهد که تابع را ساده کند تا بر روی نرخ رشد متمرکز شود؛ بنابراین توابع مختلف با نرخ رشد یکسان می‌توانند دارای یک علامت O مشابه باشند. تابع زیر بر حسب n را در نظر بگیرید

$$T(n) = ۴n^۲ - ۵n + ۷.$$

[36] Big o notation

[37] Small o notation

اگر تعداد ورودی این مساله به بی‌نهایت میل کند اندازه جمله $n^۲$ بسیار نزدیک‌تر از دیگر جمله‌ها خواهد بود. در این صورت گفته می‌شود $T(n) \in O(n^۲)$.

در ریاضیات معمولا برای نشان دادن اینکه یک سری هندسی متناهی تا چه اندازه به تابع مورد نظر نزدیک است خصوصاً در سری تیلور ناقص از این علامت استفاده می‌شود.

با این مقدمه، انتگرال رابطه (۱۷.۱) می‌تواند به‌صورت

$$\log P(\mathcal{T}|\mathcal{M}_n) = \log P\left(\mathcal{T}|\hat{\theta}_n, \mathcal{M}_n\right) - \frac{p_n}{۲}\log N + O(۱)$$

نوشته شود. بر اساس رابطه (۱۷.۱)،

$$\log P(\mathcal{M}_n|\mathcal{T}) = \log(P(\mathcal{M}_n)) + \log P\left(\mathcal{T}|\hat{\theta}_n, \mathcal{M}_n\right) - \frac{p_n}{۲}\log N.$$

اکنون یک پیشین یکنواخت روی همه مدل‌ها فرض می‌شود، اولین عبارت می‌تواند حذف شود و فقط عبارت

$$\log P\left(\mathcal{T}|\hat{\theta}_n, \mathcal{M}_n\right) - \frac{p_n}{۲}\log N.$$

باید ماکسیمم شود که این مسئله با کمینه کردن معیار شوارتز به‌صورت

$$-\log P\left(\mathcal{T}|\hat{\theta}_n, \mathcal{M}_n\right) + \frac{d_n}{۲}\log N$$

معادل است. در مدل خطی با خطاهای نرمال، این به معیار BIC کاهش می‌یابد (تعریف ۵.۲.۱).

تصمیم‌گیری در مورد استفاده از کدام یک از معیارهای BIC یا AIC در اغلب موارد، دشوار است. معیار BIC، از مزیت به‌طورمجانبی سازگار بودن ــ هر چه $N \to \infty$، BIC مدل درست را انتخاب می‌کند ــ برخوردار است. با این وجود، در نمونه‌های متناهی، معیار BIC بنا به جریمه سنگینی که بر مدل‌های پیچیده می‌کند، مدل‌هایی که بسیار کوچک هستند را انتخاب می‌کند.

تحلیل کلی

معیارهای AIC و BIC با در نظر گرفتن خطای مدل‌ساز و تصحیح آن، یک برآورد از خطای کلی موردانتظار ارائه می‌دهند. این‌ها فقط دو تا از تعداد زیادی برآوردگرهایی از این نوع هستند (برای مثال، هوکینگ، ۱۹۷۶ را ببینید). هدف مشترک بیشتر این برآوردگرها، کمینه کردن کمیتی است که نسبت به $\overline{\text{err}}$ اکیدا صعودی است. به‌عنوان مثال، معیارهای AIC و BIC هر دو عبارت

$$Q(\boldsymbol{\beta}) = \overline{\text{err}} + \lambda||\boldsymbol{\beta}||_۰ \qquad (۱۸.۱)$$

$$= \frac{۱}{N}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^۲ + \lambda||\boldsymbol{\beta}||_۰ \qquad (۱۹.۱)$$

را کمینه می‌کنند (که در آن $p = ||\boldsymbol{\beta}||_۰$ به مفهوم نرم $L_۰$ بردار $\boldsymbol{\beta}$ است؛ به‌عبارت دیگر، تعداد مولفه‌های غیرصفر در بردار $\boldsymbol{\beta}$ را بیان می‌کند). درباره C_p و AIC، $\lambda = ۲\sigma_\epsilon^۲/N$ و برای BIC، $\lambda = \sigma_\epsilon^۲ \log N/N$.

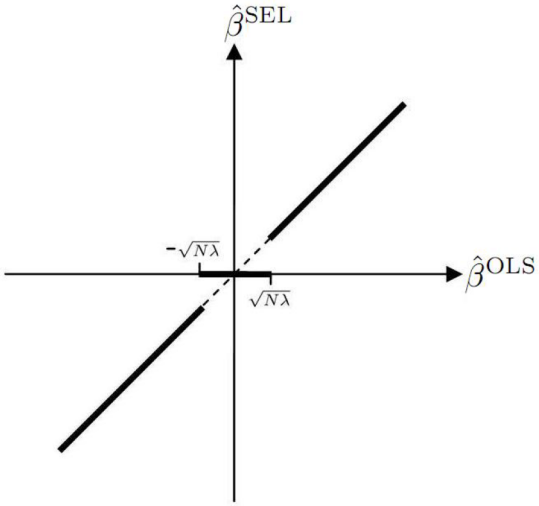
برای درک بهتر، این هدف در حالت متعامد بررسی می‌شود. بر اساس قضیه ۳.۲.۱ و بحث‌های پس از آن، وقتی $\mathbf{X}^\top\mathbf{X} = \boldsymbol{I}$، مسئله کمینه کردن رابطه (۱۹.۱) هم‌ارز با کمینه کردن مسئله

$$q(\beta_j) = \frac{1}{N}\left(\hat{\beta}_j^{\mathrm{OLS}} - \beta_j\right)^2 + \lambda \mathbb{1}_{\beta_j \neq 0}$$

برای هر β_j به‌طور جداگانه است. یک پاسخ برای این مسئله کمینه‌سازی برای هر β_j به‌وضوح عبارت است از

$$\hat{\beta}_j = \hat{\beta}_j^{\mathrm{OLS}} \mathbb{1}_{|\hat{\beta}_j^{\mathrm{OLS}}| > \sqrt{N\lambda}}.$$

به‌عبارت دیگر، این روش برآورد کمترین توان‌های دوم معمولی $\boldsymbol{\beta}$ را پیدا می‌کند و هر مولفه برآورد را درنظر می‌گیرد. هر مولفه با قدر مطلق مقدار بیشتر از $\sqrt{N\lambda}$ منقبض نمی‌شود. بقیه مولفه‌ها به‌سمت صفر منقبض می‌شوند. این رفتار در شکل ۳.۱ نشان داده شده است. درنظر گرفتن این روش بر حسب مبادله اریبی–واریانس، حاوی اطلاعات مفیدی است. با منقبض کردن برخی مولفه‌های $\boldsymbol{\beta}$، مقداری اریبی ایجاد می‌شود – زیرا برآورد نااریب کمترین توان‌های دوم به‌وضوح نشان می‌دهد که این مولفه‌ها صفر نیستند. با این وجود، با صفر قرار دادن این مولفه‌ها، واریانس آن‌ها کاهش پیدا می‌کند – و از این رو، دقت پیشگویی در آن روش افزایش می‌یابد.



شکل ۳.۱: رفتار مدل‌هایی مانند C_p، AIC و BIC. برآوردهای کمترین توان‌های دوم معمولی محاسبه شده است و هر مولفه کوچک آن به سمت صفر منقبض شده است.

اگر β_j خیلی بزرگ باشد، انقباض آن به سمت صفر منجر به افزایش خیلی زیادی در اریبی می‌شود ــ کاهش واریانسی که در نتیجه ایجاد اریبی به‌وجود می‌آید، احتمالا برای جبران این مسئله کافی نیست. اگر از طرف دیگر، β_j کوچک باشد، کاهش در واریانس احتمالا افزایش در اریبی را جبران می‌کند. روش‌های مطرح شده نقطه «برش[۳۸]» را در $\beta_j = \sqrt{N\lambda}$ قرار می‌دهند. در معیارهای C_p و AIC، نقطه برش، $\beta_j = ۲\sigma_\epsilon^۲$ و در معیار BIC، $\beta_j = \sigma_\epsilon^۲ \log N$ است.

کمترین توان‌های دوم جریمه شده صورت رابطه (۱۸.۱)، خانواده کلی‌تری از روش‌هایی را پیشنهاد می‌دهد که کمیت

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \overline{\text{err}} + \sum_{j=۱}^{p} p_\lambda(|\beta_j|) \\ &= \frac{۱}{N}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^۲ + \sum_{j=۱}^{p} p_\lambda(|\beta_j|) \end{aligned} \qquad (۲۰.۱)$$

را کمینه می‌کند که در آن $P_\lambda(.)$ تابع جریمه نامیده می‌شود. وابستگی تابع جریمه به پارامتر تنظیم‌کننده[۳۹] λ، روشی برای شدت جریمه برای نیشگون[۴۰] یا پیچاندن پارامترها را نشان می‌دهد.

در این نقطه کنونی، به نظر نمی‌رسد کمترین توان‌های دوم جریمه‌شده یک رهیافت مطلوب باشد. برخلاف رهیافت‌های دیگری که تاکنون بحث شده، به نظر می‌رسد هیچ انگیزه تئوری خوبی برای استفاده از آن وجود ندارد. علی‌رغم این عیب، روش کمترین توان‌های دوم جریمه‌شده، یکی از موفق‌ترین روش‌ها برای انتخاب مدل بعد بالا است.

[38] Cut-off
[39] Regularization parameter
[40] tweak

۶.۲.۱ مسئله بعد بالا

برای درک لزوم استفاده از روش کمترین توان‌های دوم جریمه‌شده، باید مسیر به‌کارگیری روش‌هایی مانند C_p و BIC . پیموده شود. هر یک از این معیارها، یک برآورد از Err هستند که تلاش می‌شود کمینه شود و همه به شکل

$$\overline{\text{err}} + \lambda P_\lambda$$

هستند. این بدین معناست که وقتی مشخص می‌شود از کدام متغیرها در مدل استفاده شود (به‌عبارت دیگر، در β غیرصفر باشند)، بعد p ثابت می‌شود و فقط باید با پیدا کردن جواب کمترین توان‌های دوم معمولی از معادله (۷.۱)، $\overline{\text{err}}$ کمینه شود. این موضوع روش‌های زیر را پیشنهاد می‌دهد:

- هر ترکیب ممکن از متغیرها در β را بسازید.
- برای هر یک از این ترکیب‌ها، برآورد کمترین توان‌های دوم معمولی را پیدا کنید، $\overline{\text{err}}$ و برآورد مناسب Err را به‌دست آورید.
- ترکیبی با کمترین برآورد Err را انتخاب کنید.

البته مشکل این روش این است که وقتی مدل، بعد بالا است اولین گام یک هزینه محاسباتی سنگینی را تحمیل می‌کند (اگر بردارهای ورودی اصلی $\mathbf{X}$، p مولفه داشته باشد، $۲^p - ۱$ ترکیب ممکن از متغیرها وجود دارد!). این موضوع باعث می‌شود که این روش‌ها، به‌شدت غیرکاربردی شوند. مطالعات قابل توجهی برای کاهش این هزینه محاسباتی در دو دسته زیر انجام شده‌اند.

- کاهش زمان محاسبه برآورد کمترین توان‌های دوم معمولی برای هر مدل با به‌روزرسانی ماتریس هَت $\boldsymbol{H}$ به‌جای محاسبه دوباره آن برای هر مدل و با حذف مرتب داده‌های اضافی از این ماتریس‌ها. برای جزئیات بیشتر فورنیوال (۱۹۷۱) را ببینید.
- یافتن روش‌های هوشمندانه برای حذف برخی ترکیبات به‌عنوان نشدنی بدون آزمون کردن آن‌ها. بیشتر این روش‌ها بر این حقیقت متکی هستند که کاهش $\overline{\text{err}}$ با حذف متغیرها از مدل غیرممکن است (درواقع، شکل ۱.۱ به‌وضوح نشان می‌دهد که $\overline{\text{err}}$ با افزایش p اکیدا نزولی است).

  این مفهوم با یک مثال بهتر نشان داده می‌شود. تصور کنید، هدف یک آماردان، کمینه کردن کمیت

  $$Q = \overline{\text{err}} + p$$

  است. هم‌چنین باز هم تصور کنید یک ترکیب یک متغیره $\mathcal{A}$ با $Q_{\mathcal{A}} = ۲$ وجود دارد. اکنون با یک ترکیب سه متغیره $\mathcal{B}$ با $Q_{\mathcal{B}} = ۵$ پیدا می‌شود. این روش، این اجازه را

می‌دهد همه ۷ زیرمجموعه‌های $\mathcal{B}$ نادیده گرفته شود زیرا حتی اگر با در نظر گرفتن این زیرمجموعه‌ها، $\overline{\text{err}}$ افزایش پیدا نکند، بهترین چیزی که می‌توان به‌دست آورد، برابر $Q = Q_{\mathcal{B}} - ۲ = ۳$ است که بر $Q_{\mathcal{A}} = ۲$ برتری ندارد. از این رو، فضای جستجو به ۷ کاهش پیدا می‌کند. چنین روش‌هایی مثالی از روش‌های شاخه و کران[۴۱] هستند. البته این مثال بسیار ساده‌گرایانه است. برای بحث دقیق‌تر، لاموته و هوکینگ (۱۹۷۰) را ببینید.

این دو روش در رهیافت «پرش‌ها و کران‌ها[۴۲]» (فورنیوال و ویلسون، ۱۹۷۴) ترکیب شده‌اند. این رهیافت، قادر است با ۳۰ تا ۴۰ متغیر روبرو شود اما در مقایسه با الزامات مسائل بعد بالای جدید که اغلب شامل هزاران متغیر می‌شوند، بی‌اهمیت است.

برخی الگوریتم‌های ابتکاری نیز برای کمک به این روند وجود دارند. برای مثال،

- انتخاب پیشرو[۴۳] یک الگوریتم حریص و پرخور است که از یک مدل خالی شروع می‌کند و به‌طور پی در پی متغیری که بیشترین همبستگی با پاسخ را دارد، به آن اضافه می‌شود.
- حذف پسرو[۴۴] با یک مدل از همه متغیرها شروع می‌کند و از متغیر با کمترین معنی‌داری یکی پس از دیگری حذف می‌کند. این مدل در حالتی‌که تعداد متغیرها بیشتر از N است، که در آن هیچ برآورد کمترین توان‌های دوم معمولی وجود ندارد، قابل اجرا نیست.
- انتخاب گام به گام[۴۵] اساسا انتخاب پیشرو است، اما امکان حذف تک متغیر ــ سبک انتخاب پسرو ــ در هر گام پیشرو وجود دارد. این رویه ابتکاری اخیرا فرمول‌بندی شده است و به‌طور کامل توسط ژانگ (۲۰۰۸) تحت عنوان FOBA تحلیل شده است و رقیبی برای سایر روش‌های مطرح شده است.

البته این روش‌ها، همه به یک قاعده توقف[۴۶] نیاز دارند.

با در نظر گرفتن این روش‌ها، انتخاب مدل کلاسیک یک مسئله خیلی سخت از دیدگاه محاسباتی را بدون پاسخ می‌گذارد به‌خصوص، برای مدل‌هایی که در بعدهای بالاتر قرار می‌گیرند. چگونه کمترین توان‌های دوم جریمه‌شده می‌تواند کمک کند؟ پاسخ این است که وقتی $P_\lambda(.)$ یک تابع محدب است، مسئله کمینه کردن رابطه (۲۰.۱)، یک مسئله بهینه‌سازی محدب می‌شود که دقیقا در مرتبه چندجمله‌ای قابل حل است (برای مثال، بوید و واندنبرگ، ۲۰۰۴ را برای آشنایی با تئوری بهینه‌سازی محدب ببینید). این عبارت‌ها از لحاظ نظری به جذابیت معیارهایی مانند AIC نیستند، اما حداقل قابل اندازه‌گیری هستند و اغلب می‌توانند منجر به جواب‌های خوب قابل توجهی شوند.

[41] Branch and bound
[42] Leaps and bounds
[43] Forward selection
[44] Backwards elimination
[45] Stepwise selection
[46] Stopping rule

فصل ۲

روش‌های انتخاب متغیر کلاسیک

سوالی که برای برخی محققان مطرح می‌شود این است که در برازش مدل‌های رگرسیونی، برآوردگر کمترین توان‌های دوم خوب رفتار می‌کند اما چرا این برآوردگر برای انتخاب مدل، مناسب نیست؟ در پاسخ به این سوال، می‌توان گفت یک مدل با تعداد متغیرهای زیاد، همیشه مجموع توان دوم باقی‌مانده‌های کوچکتری دارد زیرا پیدا کردن کمینه روی مجموعه بزرگتر، همیشه کوچکتر خواهد بود. بنابراین، اگر «کمترین توان‌های دوم» به‌عنوان یک شرط برای هدف انتخاب مدل درنظر گرفته شود، آن‌گاه می‌گوید «همیشه مدل بزرگتر را انتخاب کن».

همان‌طور که در فصل قبل اشاره شد، معمولا محققان علاقه‌مند به مدل‌هایی هستند که پیچیدگی کمتری داشته باشند. به‌عبارت دیگر، تعداد متغیرهای آن‌ها کمتر باشد. چنین مدل‌هایی از نظر تفسیر پذیری و بررسی روابط علت و معلولی بهتر هستند. از این نظر، بحث انتخاب متغیر و امساک[۱] یا به‌عبارت دیگر اصل تُنُک[۲] بودن جزو اهداف اصلی مدل‌سازی است. به‌علاوه باید توجه داشت که انتخاب متغیر معمولا باعث بالا رفتن دقت پیش‌بینی نیز می‌شود. به‌عبارت دیگر، وجود تعداد متغیرهای خیلی زیاد در مدل باعث بیش‌برازشی[۳] و وجود تعداد متغیرهای خیلی کم در مدل باعث کم‌برازشی[۴] می‌شود.

[1] Parsimony principle
[2] Sparsity
[3] Overfitting
[4] Underfitting

به‌طور کلی، در رگرسیون چندگانه[۵] دو سوال مهم باید پرسیده شوند:

- آیا همه متغیرها کمک می‌کنند که متغیر Y توضیح داده شود یا فقط یک زیرمجموعه از متغیرها، اثرگذار هستند؟

- اگر فقط یک زیرمجموعه از متغیرها، موثر باشند، چگونه می‌توان آن زیرمجموعه را انتخاب کرد؟ به‌عبارت دیگر، چگونه می‌توان متغیرهای بااهمیت را انتخاب کرد؟

در این فصل، روش‌های انتخاب متغیر کلاسیک که بیشتر در مسائل با بعد کم کاربرد دارند، مطرح می‌شود.

۱.۲ روش رگرسیون همه زیرمجموعه‌ها

یکی از روش‌های انتخاب زیرمجموعه‌ای از متغیرها، روش رگرسیون همه زیرمجموعه[۶]ها یا بهترین زیرمجموعه[۷] است. این روش همان‌طور که از نامش مشخص است، همه زیرمجموعه‌های ممکن از متغیرها را درنظر می‌گیرد و برای هر زیرمجموعه ممکن، برآوردگر کمترین توان‌های دوم را محاسبه می‌کند و سپس بهترین زیرمجموعه را بر اساس یک معیار مانند AIC، BIC. یا معیار اعتبارسنجی متقابل (که در فصل قبل به آن‌ها اشاره شد)، انتخاب می‌کند. مشکل اصلی این روش هزینه محاسباتی آن است. تعداد کل زیرمجموعه‌های یک مجموعه pتایی از متغیرها، 2^p است. بنابراین محاسباتی که باید انجام شود با اضافه شدن هر متغیر دو برابر می‌شود و به‌عنوان مثال، برای $p = 40$، بیشتر از یک بیلیون مدل وجود دارد. درواقع، روش همه زیرمجموعه‌ها، تضمین می‌کند که بهترین زیرمجموعه از متغیرها پیدا می‌شود اما هزینه محاسباتی فوق‌العاده بالایی دارد.

در عمل، استفاده از این روش حتی برای مواردی که شامل تعداد متوسطی متغیر مستقل هستند بسیار خسته‌کننده یا شاید نشدنی باشد. الگوریتم‌هایی برای سهولت استفاده از این روش ارائه شده‌اند که بعضی از آن‌ها در میلر (۲۰۰۲) یافت می‌شود.

برای غلبه بر هزینه بالای محاسباتی روش همه زیرمجموعه‌ها، یک رهیافت خودکار معرفی شده است که در بین همه زیرمجموعه‌ها جست‌وجو می‌کند و بهترین را انتخاب می‌کند. در ادامه، سه مورد از این رهیافت خودکار، توضیح داده می‌شود.

---

[5] Multiple regression

[6] All Subset Regression

[7] Best subset

۲.۲ روش انتخاب پیشرو

یکی از معروف‌ترین و پرکاربردترین روش‌های انتخاب زیرمجموعه‌ای از متغیرهای مستقل، روش انتخاب پیشرو[8] است. این روش به‌صورت گام‌به‌گام بوده و در هرگام یک متغیر را به مدل وارد می‌کند، فرض کنید $\mathbf{x}_1$، ...، $\mathbf{x}_p$ متغیرهای مستقل و y متغیر وابسته باشد. در روش انتخاب پیشرو، ابتدا مدل بدون هیچ متغیری درنظر گرفته می‌شود. سپس متغیری وارد مدل می‌شود که بیشترین کاهش را در مجموعه توان دوم مانده‌ها که به‌صورت

$$S_j = \sum_{i=1}^{n} (y_i - \beta_j x_{ij})^2$$

است، ایجاد کند. اولین متغیر، متغیری است که به ازای آن S_j، کمینه شود و باتوجه به اینکه

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} x_{ij} y_i}{\sum_{i=1}^{n} x_{ij}^2},$$

بنابراین

$$S_j = \frac{\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} x_{ij} y_i)^2}{\sum_{i=1}^{n} x_{ij}^2}.$$

از این جهت می‌توان گفت که در گام اول متغیری وارد مدل می‌شود که به ازاء آن

$$\frac{(\sum_{i=1}^{n} x_{ij} y_i)^2}{\sum_{i=1}^{n} x_{ij}^2}. \qquad (۱.۲)$$

ماکسیمم شود. اگر متغیرهای تبینی استاندارد شده و متغیر وابسته مرکزی شده باشد، رابطه (۱.۲) متناسب با توان دوم همبستگی بین متغیر y و $\mathbf{x}_j$ می‌شود. اگر اولین متغیر انتخاب شده با $\mathbf{x}_{(1)}$ نشان داده شود، برای انتخاب متغیر بعدی رابطه $\mathbf{x}_{j.(1)} = \mathbf{x}_j - \beta_{j.(1)}\mathbf{x}_{(1)}$ که در آن $\beta_{j.(1)}$ ضریب رگرسیونی کمترین توان‌های دوم $\mathbf{x}_j$ روی $\mathbf{x}_1$ است، تشکیل داده می‌شود. متغیر بعدی، متغیری است که به ازای آن رابطه (۱.۲) ماکسیمم شود، هنگامی‌که $y - \mathbf{x}_{(1)} b_{(1)}$ جایگزین y و $\mathbf{x}_{j.(1)}$ جایگزین $\mathbf{x}_j$ در رابطه (۱.۲) شود. این فرآیند ادامه می‌یابد تا جایی که همه متغیرها وارد مدل شوند یا یک معیار از قبل تعیین شده، نتیجه شود. می‌توان از یک الگوریتم ساده بر اساس همبستگی‌های جزئی نیز برای نمایش روش انتخاب پیش‌رو استفاده کرد. به‌طور خلاصه اگر متغیرها در شرایط نرمال‌سازی صدق کنند، آن‌گاه روش انتخاب پیشرو را می‌توان به‌صورت زیر بیان کرد:

۱. قرار دهید $\hat{\beta}_1 = \cdots = \hat{\beta}_p = 0$ آن‌گاه اگر مانده‌ها با r نشان داده شود، $r = y - \hat{y}$.

۲. متغیری که بیشترین همبستگی را با r دارد انتخاب می‌شود، به‌عنوان مثال، $\mathbf{x}_{(1)}$.

۳. y را روی $\mathbf{x}_{(1)}$ رگرسیون کرده و $r = y - \hat{y}$ محاسبه می‌شود.

[8] Forward

۴. از گام دوم به بعد تکرار می‌شود تا همه متغیرها وارد مدل شوند یا معیار از پیش تعیین شده اتفاق افتد.

معمولا در این روش از یک معیار برای توقف ورود متغیرها استفاده می‌شود. در اکثر نرم‌افزارهای آماری از آزمون F به‌عنوان معیاری برای توقف ورود متغیرها استفاده می‌شود. عملکرد این آزمون به این صورت است که تا وقتی که متغیر وارد شده به مدل کاهش معنی‌داری در مجموع توان‌های دوم خطا بدهد، به آن اجازه ورود به مدل داده می‌شود.

اخیرا با پیشرفت تکنولوژی، استفاده از برنامه‌های رایانه‌ای به‌ویژه نرم افزار R افزایش یافته است که در این نرم‌افزار، از معیارهای AIC و BIC به‌عنوان معیار توقف روش انتخاب پیش‌رو استفاده شده است. از این رو، بیان الگوریتم انتخاب پیش‌رو بر اساس این معیارها، خالی از فایده نیست. این الگوریتم به‌صورت زیر است:

- با مدل خالی یعنی مدلی که فقط شامل عرض از مبدا است و هیچ یک از متغیرهای پیش‌گو در آن حضور ندارند، شروع کنید.
- p رگرسیون خطی ساده را برازش دهید و به مدل خالی، متغیری را اضافه کنید که کمترین AIC یا BIC را دارد.
- به آن مدل، متغیری را اضافه کنید که دارای کمترین AIC یا BIC در بین همه مدل‌هایی با دو متغیر است.
- آن قدر ادامه دهید تا زمانی که در قاعده توقف مانند اینکه AIC یا BIC کمترین شود، صدق کند.

درباره تفاوت معیارهای AIC و BIC، می‌توان گفت که استفاده از معیار BIC، مهاجم[۹]ترین مدل و مدل کمتر را انتخاب می‌کند. معیارهای AIC و C_p مالو، به‌طور مشابهی رفتار می‌کنند. آن‌ها در مقایسه با معیار BIC کمتر به متغیرها هجوم می‌برند و بنابراین، مدل‌های بزرگتری در مقایسه با معیار BIC انتخاب می‌شوند. از این رو، اگر هدف یک محقق، انتخاب یک مدل صرفه‌جوتر است، از BIC استفاده کنید. هنگامی که از معیار AIC برای انتخاب مدل استفاده می‌شود، بهتر است مدل تولید شده نهایی را چک کنید. برخی از علامت‌های ضرایب با علامت ضرایب مشابه در مدل نهایی به‌دست آمده بر اساس معیار BIC متفاوت هستند.

انتخاب گام‌به‌گام پیشرو، یک الگوریتم حریص[۱۰] است، و مدل‌هایی از دنباله‌های تودرتو تولید می‌کند. در این منظر، این روش ممکن است در مقایسه با روش انتخاب بهترین زیرمجموعه، نامطلوب به‌نظر برسد. با این حال همیشه حتی اگر تعداد پارامترها از تعداد مشاهدات بیش‌تر باشد $(p > n)$ می‌توان آن را اجرا کرد. روش گام‌به‌گام پیشرو جستجوی

[9] Aggressive
[10] Greedy

محدودتری از نظر انتخاب بهترین زیرمجموعه برای هر اندازه‌ای نسبت به روش انتخاب بهترین زیرمجموعه است. واریانس کمتری هم خواهد داشت اما شاید اریبی را افزایش دهد.

۳.۲ روش حذف پسرو

روش انتخاب گام‌به‌گام پسرو[11] با یک مدل کامل شروع می‌شود و به‌ترتیب متغیرهایی که نسبت به پاسخ کمترین اهمیت را دارند، حذف می‌کند. نامزد حذف شدن، متغیری است که با حذف آن کمترین افزایش در مجموع توان دوم مانده‌ها رخ دهد. این روند ادامه می‌یابد تا اینکه معیاری که از قبل مشخص شده است، اتفاق بیفتد. معمولا از آزمون F به‌عنوان معیاری برای بررسی اثر خروج متغیرها از مدل استفاده می‌شود. در این صورت، خروج متغیرها از مدل تا زمانی که اثر آن‌ها در افزایش مجموع توان دوم مانده‌ها معنی‌دار نباشد، ادامه می‌یابد. انتخاب پسرو تنها می‌تواند زمانی استفاده شود که $n > p$، درحالی که روش گام‌به‌گام پیشرو همیشه قابل استفاده است.

مشابه روش انتخاب پیش‌رو، الگوریتم زیر ارائه می‌شود:

- با مدلی شروع کنید که همه متغیرها در آن است.
- هر متغیر را به‌طور جداگانه از مدل حذف کنید و AIC یا BIC نتیجه‌شده را مقایسه کنید.
- مدل $(p-1)$ـ متغیره جدید را برازش دهید و متغیری که منجر به کمترین کاهش در AIC یا BIC نسبت به مدل کامل می‌شود را از مدل حذف کنید.
- آن قدر ادامه دهید تا قاعده توقف برقرار باشد. قاعده توقف، برای مثال، می‌تواند به این صورت تعریف شود که زمانی متوقف می‌شود که همه متغیرهای باقی‌مانده، کمترین AIC یا BIC را تولید کنند.

۴.۲ روش گام به گام

این روش تا حدودی از روش‌های قبلی منطقی‌تر است. روش گام به گام ترکیبی از هر دو روش انتخاب پیشرو و حذف پسرو است. در این روش، بعد از ورود هر متغیر به مدل، متغیرهای داخل مدل برای خروج از مدل ارزیابی می‌شوند و ممکن است با ورود یک متغیر به مدل، متغیری دیگر از مدل خارج شود. این امری منطقی است زیرا همبستگی بین متغیرها می‌تواند اثر آن‌ها را تحت‌تاثیر قرار دهد. به‌عنوان مثال، متغیری را در نظر بگیرید که در مرحله سوم دارای اثر معنی‌دار بوده و وارد مدل شده است. در مراحل بعدی حضور متغیرهای دیگر در

[11] Backward

مدل و همبستگی بین آن‌ها می‌تواند اثر این متغیر را در حضور سایر متغیرها کاهش داده و باعث خروج آن از مدل شود. روند این الگوریتم به‌صورت زیر است

فرض کنید SSE_p نشان‌دهنده مجموع توان دوم مانده‌ها با حضور p متغیر و یک ثابت در مدل باشد و کمترین SSE که با حضور یک متغیر دیگر داخل مدل به‌دست می‌آید، SSE_{p+1} باشد. نسبت

$$\frac{\text{SSE}_p - \text{SSE}_{p+1}}{\text{SSE}_{p+1}/(n-p-2)} \tag{۲.۲}$$

برای مقایسه با مقدار از پیش تعیین شده F_{in} به‌کار می‌رود. اگر این نسبت از F_{in} بیشتر باشد متغیر وارد مدل می‌شود. حال فرض کنید با ورود این متغیر جدید $p+1$ متغیر و یک ثابت در مجموعه متغیرهای انتخاب شده قرار دارند و SSE_p کمترین SSE باشد که با حذف هر متغیر از مدل به‌دست می‌آید. نسبت

$$\frac{\text{SSE}_p - \text{SSE}_{p+1}}{\text{SSE}_{p+1}/(n-p-2)} \tag{۳.۲}$$

با مقدار از پیش تعیین شده F_{out} مقایسه شده و در صورتی‌که از آن کمتر باشد، متغیر مورد نظر از مدل خارج می‌شود. همگرایی این الگوریتم را نیز می‌توان نشان داد.

از رابطه (۲.۲) می‌توان نتیجه گرفت که وقتی شرط ورود یک متغیر برقرار است،

$$\text{SSE}_{p+1} \leq \frac{\text{SSE}}{\{1 + F_{\text{in}}/(n-p-2)\}}$$

و از رابطه (۳.۲) می‌توان نتیجه گرفت که متغیر خارج می‌شود اگر

$$\text{SSE}_p \leq \frac{\text{SSE}_{p+1}}{\{1 + F_{\text{out}}/(n-p-2)\}}.$$

بنابراین وقتی یک خروج بعد از یک ورود در مدل رخ می‌دهد، SSE جدید، مثلا SSE^*، در رابطه زیر صدق می‌کند

$$\text{SSE}^* \leq \text{SSE}_{p+1}\frac{1 + F_{\text{out}}/(n-p-2)}{1 + F_{\text{in}}/(n-p-2)}. \tag{۴.۲}$$

الگوریتم زمانی متوقف می‌شود که هیچ ورود یا خروجی در این شرط صدق نکند. چون هر SSE_p دارای حد پایین برابر با کمترین SSE برای هر زیرمجموعه از p متغیر است، با اطمینان از اینکه هربار که یک زیرمجموعه جدید از p متغیر حاصل می‌شود SSE کاهش می‌یابد همگرایی تضمین می‌شود. از رابطه (۴.۲) می‌توان دریافت که یک شرط کافی برای همگرایی آن است که $F_{\text{out}} < F_{\text{in}}$ باشد.

فصل ۳

روش‌های انتخاب متغیر مدرن

ما وارد عصر داده‌های بزرگ یا مه‌داده‌ها [1] شده‌ایم:

- داده‌ها می‌توانند به طور پیوسته‌ای تولید شوند و با هزینه اندکی ذخیره شوند.
- داده‌ها می توانند از ابعاد بالایی باشند یا ساختار آنها مشخص نباشد.
- روند داده‌های بزرگ تمایلی به حفظ یا حتی سرعت بخشیدن دارد.
- حجم انبوهی از داده‌های با بعد بالا چالش‌ها و شانس‌های جدیدی را برای تحلیل‌گران داده فراهم می کند.
- پیشرفت‌های علوم در زمینه استخراج از داده‌ها بیشتر و بیشتر شده اند.
- تجزیه و تحلیل‌های معتبر آماری برای داده‌های بزرگ به سرعت اهمیت یافته است.

در داده‌های با بعد بالا، تعداد متغیرها p ،خیلی بزرگ است. گاهی اوقات، آن را با رابطه $p = \exp(n^\alpha)$، $0 < \alpha < 1$، نشان می‌دهند. اگر n هم خیلی بزرگ باشد، در حالت کلی، به آن داده‌ها، مه‌داده‌ها می‌گویند. در داده‌های بعد بالا، در بین برخی پیش‌گوها، همبستگی بالایی وجود دارد.

معمولا پذیره‌های زیربرای تحلیل داده‌های با بعد بالا در یک مدل برقرار هستند:

[1] Big data

- مدل تنک است، به این معنی که برخی متغیرهای پیشگو، متغیرهای مزاحم هستند و بر متغیر پاسخ، اثر ندارند.
- تنکی، ساختار خاصی دارد.

مسئله اصلی در داده‌های بعد بالا این است که چگونه می‌توان متغیرهای رابااهمیت با ساختار تنکی خاصی پیدا کرد. بنابراین می‌توان بیان کرد که یکی از اهداف اصلی تحلیل داده‌های بعد بالا و مه‌داده‌ها، انتخاب متغیرها است. بر اساس بیکل و لونیا (۲۰۰۸)، در تحلیل داده‌ها با بعد بالا:

- باید روش‌های موثر را بهبود ببخشید به‌طوری‌که مشاهدات آینده و مشاهدات در همان زمان را بادقت مناسبی پیش‌بینی کنند.
- باید به بینشی درباره ارتباط بین متغیرهای پیش‌گو و پاسخ برای اهداف علمی دست یافت.

از روش‌هایی که می‌توان به آن‌ها در تحلیل داده‌های با بعد بالا اشاره کرد، روش‌های جریمه شده و غربال‌گری مستقل مطمئن[۲] هستند.

۱.۳ کمترین توان‌های دوم جریمه شده

در این بخش،روش کمترین توان‌های دوم جریمه‌شده که در فصل ۱ به آن اشاره شد، با جزئیات بیشتری مدنظر قرار می‌گیرد. دراین روش، $\boldsymbol{\beta}$ای انتخاب می شود که عبارت

$$\frac{1}{N}||\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}||^2+\sum_{j=1}^{d}p_\lambda(|\beta_j|) \tag{۱.۳}$$

را کمینه کند. در این بخش، طیف وسیعی از جریمه‌های p_λ در نظر گرفته می‌شود. در تحلیل آن‌ها، مورد توجه قرار دادن حالت طرح متعامد مفید خواهد بود. در مورد طرح متعامد، از قضیه ۳.۲.۱ نشان داد که کوچک کردن رابطه (۱.۳) هم‌ارز با کمینه کردن عبارت

$$q(\beta_j)=\frac{1}{N}\left(\hat{\beta}_j^{\mathrm{OLS}}-\beta_j\right)^2+p_\lambda(|\beta_j|) \tag{۲.۳}$$

برای هر β_j است. با این حال، قبل از آغاز بحث، ابتدا انتظارات یک محقق و آماردان از p_λ در نظر گرفته می‌شود.

[2] Sure Independent Screening (SIS)

۱.۱.۳ ملاحظات کلی

فن و لی (۲۰۰۱) پیشنهاد کردند که هر تابع جریمه p_λ باید منجر به یک برآوردگر $\boldsymbol{\beta}$ شود که ویژگی‌های زیر را تحقق بخشد:

تُنُکی ـ در برخی مدل‌های بعد خیلی بالا، دلایل از قبل مشخصی وجود دارند که معتقد هستند باید برخی از مولفه‌های $\boldsymbol{\beta}$ صفر باشند. از این رو، انتظار می‌رود که هر p_λ که انتخاب می‌شود منجر به روشی شود که برخی مولفه‌های $\boldsymbol{\beta}$ را برابرصفر برآوردکند.

تُنُکی نیز در کوچک کردن عدد موثر پارامترها در مدل (تعریف ۳.۲.۱) و کاهش خوش‌بینی ω سودمند، اما با اهمیت کمتر، نقش دارد. درواقع، حتی اگر پارامترها به دقیقا صفر منقبض نشوند، باز هم کاهشی در عدد موثر پارامترها اتفاق می‌افتد.

ناااریبی ـ کمترین توان‌های دوم جریمه شده و روش‌های مشابه، مقداری اریبی به $\boldsymbol{\beta}$ به امید کاهش واریانس اضافه می‌کند. هرچند این پدیده، اکثرا زمانی رخ می‌دهد که β_j کوچک است و روش پیشنهادی، آن را به سمت صفر منقبض می‌کند. از این رو، انتظار می‌رود که هر p_λی که انتخاب می‌شود منجر به برآوردگری شودکه تقریبا برای مولفه‌های بزرگ $\boldsymbol{\beta}$ نااریب است.

پیوستگی ـ برای پیش‌گیری از ناپایداری در پیش‌بینی، باید پیشگویی روش پیشنهادی در $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ پیوسته باشد (به‌عبارت دیگر، در داده‌های مجموعه مدل‌ساز).

در ادامه، هریک از این ویژگی‌ها در طرح متعامد با جزئیات مطرح را می‌کنیم. هدف در حالت متعامد، کمینه کردن رابطه (۲.۳) برای هر مولفه β_j است. اکنون فرض کنید $\Delta = ۲/N$ و $\mathrm{sgn}(x)$ تابع علامت x باشد. بنابراین، مشتق معادله (۲.۳) نسبت به β_j به‌صورت

$$\begin{aligned} q'(\beta_j) &= -\Delta\hat{\beta}_j^{\mathrm{OLS}} + \Delta\beta_j + \mathrm{sgn}(\beta_j)p'_\lambda(|\beta_j|) \\ &= \mathrm{sgn}(\beta_j)\{\Delta|\beta_j| + p'_\lambda(|\beta_j|)\} - \Delta\hat{\beta}_j^{\mathrm{OLS}} \end{aligned}$$

است.

دقت کنید که مشتق فقط زمانی وجود دارد که تابع جریمه p_λ در همه نقاط مشتق‌پذیر باشد. در غیر این صورت، بایداز روش‌های زیرمشتق‌ها[۳] استفاده کرد.

کمینه کردن $q(\beta_j)$ هم‌ارز با پیدا کردن نقطه‌ای است که به‌ازای آن، $q'(\beta_j) = ۰$. در این راستا، فرض کنید هدف یک محقق یافتن مقدار $\boldsymbol{\beta}$ ای باشد که تابع

$$q(\boldsymbol{\beta}) = \frac{۱}{N}||\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}||^۲ + p'_\lambda(|\boldsymbol{\beta}|)$$

را هنگامی‌که تابع p_λ در صفر، مشتق‌پذیر نباشد،کمینه کند.بدین منظور، به‌مفهوم زیرمشتق‌ها نیاز است.

[3] Subgradients

تعریف ۱.۱.۳ (زیرمشتق‌ها)**.** فرض کنید $f : \mathbb{R}^p \to (-\infty, +\infty)$ یک تابع محدب باشد. نقطه $\mathbf{x}^* \in \mathbb{R}^p$ یک زیرمشتق از f در $\mathbf{x}$ است ،اگر

$$f(\boldsymbol{y}) \geq f(\mathbf{x}) + < \mathbf{x}^*, \boldsymbol{y} - \mathbf{x} > \qquad \forall \boldsymbol{y} \in \mathbb{R}^p$$

به این مفهوم که $\mathbf{x}$ مشتق صفحه‌ای است که در نقطه $\boldsymbol{y}$ با تابع f برخورد می‌کند و سایر نقاط زیر آن قرار می‌گیرند.

در نقطه‌ای که f پیوسته است، یک زیرمشتق تکی وجود دارد که با مشتق آن برابر است. در نقطه دیگر، مجموعه‌ای از زیرمشتق‌ها وجود دارند که آن‌ها را با نماد

$$\partial f(\mathbf{x})$$

نشان می‌دهند. دقت کنید که

$$\boldsymbol{o} \in \partial f(\mathbf{x}) \iff f(\boldsymbol{y}) \geq f(\mathbf{x}) \quad \forall \boldsymbol{y} \in \mathbb{R}$$

به‌عبارت دیگر، $\mathbf{x}$ کمینه تابع f است اگر و فقط اگر $\circ$ در مجموعه زیرمشتق‌های آن در آن نقطه وجود داشته باشد.

اکنون با استفاده از این تعریف، $q(\beta)$ به‌آسانی کمینه می‌شود.

تُنُکی

برای تُنُکی، روش پیشنهادی باید حداقل توانایی قرار دادن بعضی از β_j هارا برابرصفر داشته باشد. به‌عبارت دیگر، تنکی مستلزم وجود بعضی حالت‌هایی است که $q'(\circ) = \circ$. اکنون موقعیتی را درنظر بگیرید که

$$\Delta|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}| < \inf_{\boldsymbol{\beta}\in\mathbb{R}} \{\Delta|\boldsymbol{\beta}| + p'_\lambda(|\boldsymbol{\beta}|)\}. \qquad (۳.۳)$$

در این حالت، $q'(\boldsymbol{\beta})$ برای $\boldsymbol{\beta} > \circ$، مثبت و برای $\boldsymbol{\beta} < \circ$، منفی است و از این رو، $q'(\circ) = \circ$.

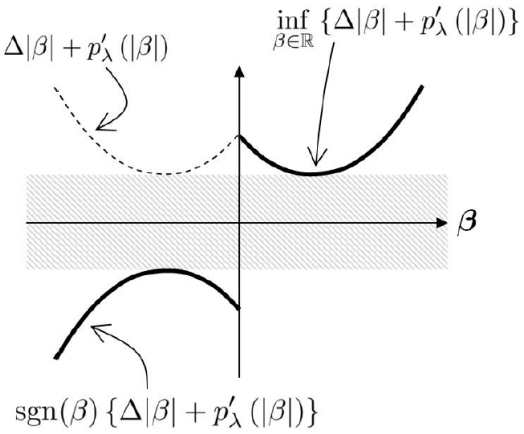
برای اینکه شرط (۳.۳) قابل اجرا باشد، باید اینفیمم سمت راست، مثبت باشد. بنابراین، شرط تُنُکی عبارت است از

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}} \{\Delta|\boldsymbol{\beta}| + p'_\lambda(|\boldsymbol{\beta}|)\} > \circ. \qquad (۴.۳)$$

این مفهوم در شکل ۱.۳ نمایش داده شده است.

ناااریبی

برآوردگر کمترین توان‌های دوم معمولی $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ نااریب است. بنابراین، به‌سادگی باید برآوردگر پیشنهادی برای مقادیر بزرگ $\boldsymbol{\beta}$ برابر با $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ باشد. اگر $p'_\lambda(|\boldsymbol{\beta}|) = \circ$، آن‌گاه $q'(\boldsymbol{\beta}) = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$



شکل ۱.۳: یک مثال از شرط تُنُکی. منحنی پررنگ تابع $\mathrm{sgn}(\boldsymbol{\beta})\{\Delta|\boldsymbol{\beta}|+p'_\lambda(|\boldsymbol{\beta}|)\}$ است و منحنی خط چین، تابع $\Delta|\boldsymbol{\beta}|+p'_\lambda(|\boldsymbol{\beta}|)$ است. اگر $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ در منطقه هاشور زده خاکستری بالا قرار بگیرد، واضح است که $q'(\boldsymbol{\beta})=-|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}|$ منحنی پررنگ است. برای $\boldsymbol{\beta}$ مثبت، مثبت خواهد بود و برای $\boldsymbol{\beta}$ منفی، منفی خواهد بود. با این حال اگر کمینه $\Delta|\boldsymbol{\beta}|+p'_\lambda(|\boldsymbol{\beta}|)$ (خط چین) مثبت نباشد (به‌عبارت دیگر، شرط (۴.۳) برقرار نباشد)، منطقه خاکستری وجود ندارد تا اتفاق بیافتد.

و از این رو، به‌وضوح در $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ یک ریشه دارد. بنابراین، شرط ناریبی این است که برای مقادیر بزرگ $\boldsymbol{\beta}$، همان برآوردهای OLS به‌دست آید. لذا به‌طور خلاصه می‌توان گفت شرط ناریبی این است که اگر $\boldsymbol{\beta}$ بزرگ شود یا به‌طور دقیق‌تر $|\boldsymbol{\beta}|\to\infty$ داشته باشیم

$$p'_\lambda(|\boldsymbol{\beta}|)\to \circ \qquad (۵.۳)$$

پیوستگی

برای اینکه برآوردگر $\boldsymbol{\beta}$ پیوسته باشد، هیچ‌وقت نباید با تغییر داده ورودی، از یک مقدار به مقدار دیگر به‌طور ناپیوسته پرش کند.

اگر شرط (۳.۳) محقق شده باشد، $\boldsymbol{\beta}$ به سمت صفر منقبض می‌شود. از این رو، حالت دیگری را به‌صورت

$$\Delta|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}|>\inf_{\boldsymbol{\beta}\in\mathbb{R}}\{\Delta|\boldsymbol{\beta}|+p'_\lambda(|\boldsymbol{\beta}|)\}$$

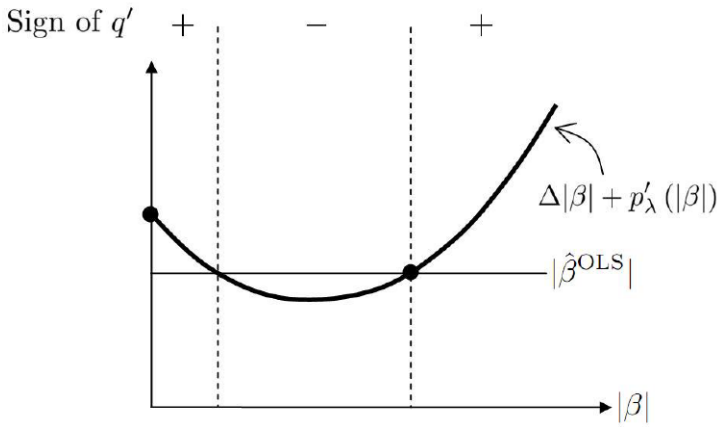
در نظر بگیرید. در این موقعیت یک شرط ضروری و کافی برای پیوستگی این است که کمینه $\Delta|\boldsymbol{\beta}|+p'_\lambda(|\boldsymbol{\beta}|)$، عدد صفر باشد. به‌عبارت دیگر،

$$\arg\inf_{\boldsymbol{\beta}\in\mathbb{R}}\{\Delta|\boldsymbol{\beta}|+p'_\lambda(|\boldsymbol{\beta}|)\}=\circ \qquad (۶.۳)$$

این شرط تقریبا کافی است، زیرا برای اینکه تابع $\Delta|\boldsymbol{\beta}|+p'_\lambda(|\boldsymbol{\beta}|)$ برای هر $\boldsymbol{\beta}\in\mathbb{R}$، تک‌مدی[۴] باشد باید t^*ای وجود داشته باشد که تابع به‌ازای هر $t<t^*$، نزولی و برای هر $t>t^*$، صعودی است.

[4] unimodal

شکل ۲.۳ نشان می‌دهد که چرا وقتی این حالت برقرار نیست، پیوستگی وجود ندارد و شکل ؟؟ توضیح می‌دهد چگونه این شرط همه چیز را درست می‌کند. برای جمع‌بندی ویژگی‌های



شکل ۲.۳: نمایشی از آن‌چه اتفاق می‌افتد وقتی شرط پیوستگی (رابطه (۶.۳)) برقرار نباشد. خط پررنگ، $\Delta|\boldsymbol{\beta}| + p'_\lambda(|\boldsymbol{\beta}|)$ و خط معمولی $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ است. بنابراین خط معمولی – خط پررنگ $= q'(\boldsymbol{\beta})$. علامت $q'(\boldsymbol{\beta})$ در نقاط مختلف در منحنی مشخص شده است و به‌وضوح نشان می‌دهد که $q(\boldsymbol{\beta})$ ابتدا افزایش، سپس کاهش و سپس دوباره افزایش می‌یابد. این بدان معناست که کمینه q باید در یکی از نقاط مشخص شده با نقطه توپر مشکی اتفاق بیافتد. با این حال، در کدام یک از آن دو نقطه باید اتفاق بیفتند، به این بستگی دارد که $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ کجاست و از این رو، نقطه‌ای که در آن پاسخ به‌طور ناپیوسته از یک نقطه به دیگری پرش می‌کند بستگی دارد.

کلی تابع جریمه p_λ، پرسیدن اینکه آیا این ویژگی‌ها برای رهیافت‌های انتخاب مدل کلاسیک (مثل AIC و BIC) برقرار هستند، آموزنده است. در این حالت،

$$p_\lambda(|\beta_j|) = \lambda \mathbb{1}_{(|\beta_j| \neq 0)}.$$

دقت کنید که مشتق p_λ، تابع دلتا دیراک[5] به‌صورت

$$p'_\lambda(|\beta_j|) = \lambda \delta(|\beta_j|)$$

است. اکنون سه ویژگی، در نظر گرفته می‌شود:

تُنُکی شرط تُنُکی در این حالت عبارت است از

$$\inf_{\boldsymbol{\beta} \in \mathbb{R}} \{\Delta|\boldsymbol{\beta}| + p'_\lambda(|\boldsymbol{\beta}|)\} > 0$$

بدیهی است که این اینفیمم نمی‌تواند منفی باشد، زیرا هردو مولفه مثبت هستند. علاوه بر این، در $\beta = 0$ تابع به‌وضوح صفر نیست زیرا تابع دلتا دیراک در این نقطه یک قله دارد. بنابراین اینفیمم بزرگتر از صفر است و شرط تحقق یافته است.

ناارِیبی به‌وضوح، برای $|\beta|$ بزرگ، تابع دلتا دیراک صفر می‌شود. بنابراین شرط ناارِیبی تحقق یافته است.

[5] Delta dirac function

پیوستگی شرط پیوستگی، عبارت است از

$$\arg\inf_{\beta\in\mathbb{R}}\{\Delta|\boldsymbol{\beta}| + p'_\lambda(|\boldsymbol{\beta}|)\} = ۰$$

متاسفانه واضح است که شرط پیوستگی برقرار نیست ــ به‌سختی می‌توان ادعا کرد که یک کمینه‌کننده در $\beta = ۰$ رخ می‌دهد که تابع دلتا دیراک ثابت می‌شود.

از این رو، انتخاب مدل کلاسیک منجر به جواب‌هایی می‌شود که تُنُک و ناریب هستند اما باتوجه به شرایط ورودی ناپایدار هستند. بحث گسترده این موضوع را در برایمن (۱۹۹۶) ببینید. بنابراین، به‌وضوح این مزیت دیگری از کمترین توان‌های دوم جریمه‌شده بر انتخاب مدل کلاسیک است. این مدل‌ها، کاربردهای بسیاری دارند. در نوروزی‌راد (۱۳۹۶) برآوردگرهای جریمه‌شده در برخی از مدل‌های رگرسیونی بررسی شده اند.

در ادامه صورت‌های مختلفی از p_λ بررسی می‌شوند.

۲.۱.۳ رگرسیون بریج

دو تا از مهم‌ترین توابع جریمه که مورد توجه قرار می‌گیرند، جریمه ریج و جریمه لاسو، حالت‌های خاصی از جریمه رگرسیون بریج (فرانک و فریدمن، ۱۹۹۳) به‌صورت

$$p_\lambda(\beta_j) = \lambda|\beta_j|^\gamma \qquad \gamma \neq ۰ \tag{۷.۳}$$

هستند. حالت $\gamma = ۰$ متناظر با انتخاب مدل کلاسیک است. از آن‌جایی که معمولا $۰^۰ = ۱$، رابطه (۷.۳) شامل این مورد خاص نمی‌شود. برای مطالعه بیشتر درباره رگرسیون بریج، خواننده علاقه‌مند به آرست (۱۳۹۵) و گودرزی (۱۳۹۷) ارجاع داده می‌شود.

در این بخش، تعدادی از ویژگی‌های کلی رگرسیون بریج در نظر گرفته می‌شود. سپس برآوردگرهای ریج و لاسو با جزئیات بیشتری بررسی می‌شوند.

شرط‌ها برای p_λ

ابتدا بررسی می‌شود که رگرسیون بریج در کدام‌یک از ویژگی تابع جریمه صدق می‌کند. در همه حالت‌ها، برای $\gamma \neq ۰$، داریم

$$p'_\lambda(\beta_j) = \lambda\gamma|\beta_j|^{\gamma-۱}$$

(مورد $\gamma = ۰$ متناظر با انتخاب مدل کلاسیک است).

تُنُکی شرط تُنُکی در این حالت عبارت است از

$$\inf_{\beta\in\mathbb{R}}\{\Delta|\beta| + p'_\lambda(|\beta_j|)\} > ۰.$$

بنابراین

- برای $\gamma > ۱$ تابع در $\beta = ۰$ به‌خوبی تعریف می‌شود و بنابراین اینفیمم در $\beta = ۰$، مقدار صفر می‌پذیرد. پس تُنُکی محقق نشده است.
- برای $\gamma \leq ۱$، هنگامی‌که $\beta \to ۰$، تابع بسیار بزرگ می‌شود و از این رو، اینفیمم بزرگتر از صفر است. بنابراین، تُنُکی محقق شده است.

ناایبی برای مقادیر بزرگ $|\beta|$، p'_λ به‌ازای $\gamma < ۱$، صفر می‌شود. بنابراین، جریمه بریج تنها به‌ازای $\gamma < ۱$، ناایب است.

پیوستگی شرط پیوستگی در این حالت عبارت است از

$$\arg \inf_{\beta \in \mathbb{R}} \{\Delta|\beta| + p'_\lambda(|\beta_j|) = ۰ \qquad (۸.۳)$$

پس

- برای $\gamma > ۱$ تابع مقدار صفر در نقطه $\beta = ۰$ می‌گیرد که به‌وضوح، اینفیمم است. بنابراین، پیوستگی محقق شده است.
- برای $\gamma = ۱$، تابع p_λ مشتق‌پذیر نیست و باید از روش زیرمشتق‌ها استفاده شود. زیرمشتق‌های β عبارت هستند از

$$\partial|\beta| = \begin{cases} -۱ & \beta < ۰ \\ \{\theta : \theta \in [-۱, ۱]\} & \beta = ۰ \\ ۱ & \beta > ۰ \end{cases}$$

به‌وضوح، کمترین مقدار زیرمشتق (-۱) می‌تواند در نقطه $\beta = ۰$ به‌دست آید. بنابراین، اینفیمم در حقیقت، در $\beta = ۰$ اتفاق می‌افتد.

- برای $\gamma < ۱$، p_λ نه مشتق‌پذیر و نه محدب است، بنابراین، حتی مفهوم زیرمشتق‌ها کمکی نمی‌کند. با این حال، در عمل، مشخص شده است که روش‌های با $\gamma < ۱$ پیوسته نیستند. این به‌طور شهودی قابل درک است ــ یک نگاه دقیق در رابطه (۸.۳) مشخص می‌کند که وقتی $\beta \to ۰$، مشتق جریمه به بی‌نهایت میل می‌کند. از این رو، بعید به‌نظر می‌رسد که رابطه (۸.۳) در $\beta = ۰$ بتواند کمینه شود.

جنبه مهم دیگری که باید در نظر گرفته شود، محدب بودن تابع جریمه است. مسئله‌ای که شامل یک تابع جریمه محدب شود، به‌آسانی می‌تواند بهینه شود. حل سایر مسائل، بسیار سخت‌تر هستند. در این حالت، تنها توابع جریمه با $\gamma \geq ۱$ محدب هستند.

این اطلاعات در جدول ۱.۳ خلاصه شده‌اند. به‌وضوح هیچ تابع جریمه‌ای به تنهایی در همه ویژگی‌ها صدق نمی‌کند. برآوردگر لاسو مزیت محدب، پیوسته و تُنُک بودن را دارد.

جدول ۱.۳: ویژگی‌های توابع جریمه

	انتخاب مدل کلاسیک	رگرسیون بریج (لاسو)		(ریج)	الاستیک‌نت	SCAD
		$\gamma < 1$	$\gamma = 1$	$\gamma > 1$		
تنکی	✓	✓	✓		✓	✓
نااریبی	✓	✓				✓
پیوستگی			✓	✓	✓	✓
محدب بودن			✓	✓	✓	

نمودارهای جریمه

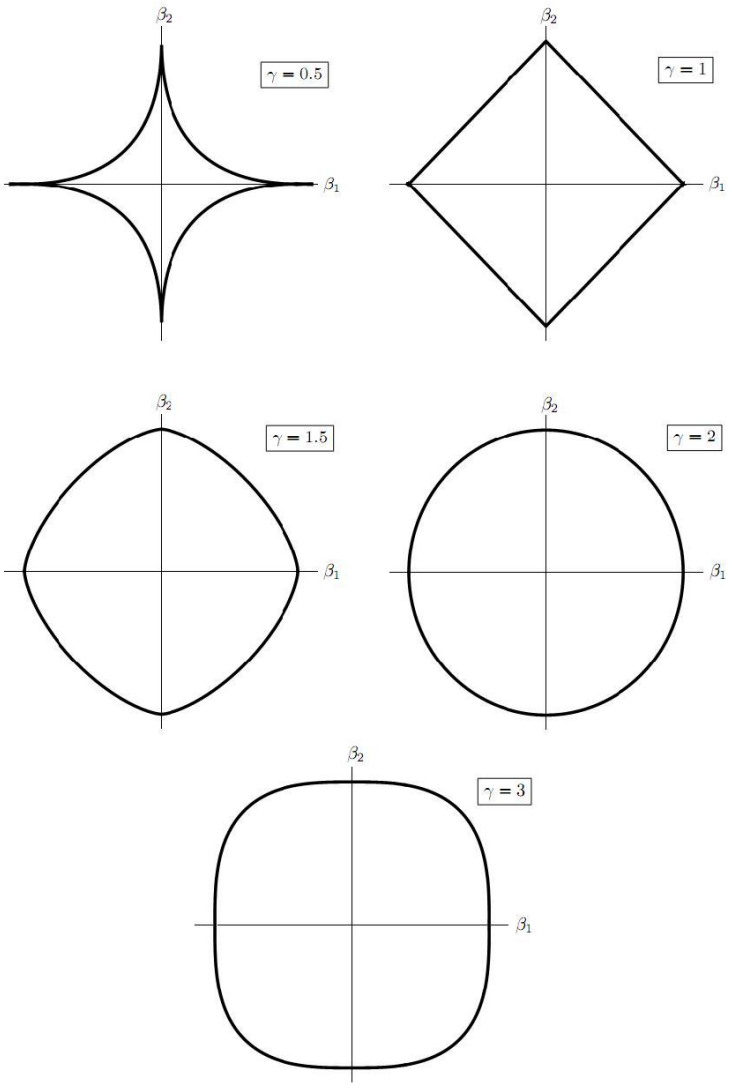
بهتراست یک محقق قبل از اینکه وارد دنیای ریاضیات رگرسیون بریج شود، شکل توابع جریمه بریج (شکل ۳.۳) برای مقادیر مختلف γ را بررسی کند. این شکل‌ها در توضیح دادن رفتار جریمه‌های بریج بسیار مفید هستند.

- برای $\gamma \leq 1$، بدیهی است که جریمه‌ها به جهتی در راستای محورهای مختصات بیشتر از سایر جهت‌ها کشش دارند ــ به‌عبارت دیگر، جریمه‌ها به جهت‌هایی که یکی از مختصات آن‌ها کوچک است، کشیده می‌شوند. بنابراین، جریمه‌ها منجر به بردارهای تُنُک β می‌شوند.

- برای $1 < \gamma < 2$ جریمه‌ها هنوز تمایل به هم جهتی در راستای محورهای مختصات را دارند. به‌علاوه توزیع‌ها در محورهای مختصات دارای گوشه نیستند و بنابراین جواب‌های تُنُک را نتیجه نمی‌دهند.

- برای $\gamma = 2$، جریمه به هیچ جهت خاصی کشش ندارد. البته انقباض هنوز اتفاق می‌افتد، زیرا نرم بردار β کوچک شده است اما در جهت مشخصی نیست.

- برای $\gamma > 2$ جریمه به جهت‌های دور از محورهای مختصات کشش دارد؛ به‌عبارت دیگر، علاقه‌اش به پارامترهای بزرگ است! به‌وضوح، این جریمه در زمینه انقباضی مطلوب نیست.



شکل ۳.۳: جریمه‌های بریج برای مقادیر مختلف γ، در حالتی که بردار $\boldsymbol{\beta} = (\beta_1, \beta_2)$. منحنی ترسیم‌شده یکی از منحنی‌های تراز $|\beta_1|^\gamma + |\beta_2|^\gamma$ هستند.

رهیافت بیزی

این که به رگرسیون بریج از یک چشم‌انداز بیزی نگاه شود، نیز بسیار آگاهی‌بخش است.
کمینه‌سازی جریمه بریج همچنین می‌تواند به‌عنوان ماکسیمم‌سازی لگاریتم توزیع پسین $(\boldsymbol{\beta}|\mathbf{Y})$ درنظر گرفته شود که به‌صورت

$$(\boldsymbol{\beta}|\mathbf{Y}) \sim C \exp\Big(-\frac{۱}{N}||\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}||^۲ - \lambda' \sum |\beta_j|^\gamma \Big)$$

است. بسیار مناسب‌تر خواهد بود اگر به‌صورت

$$(\boldsymbol{\beta}|\mathbf{Y}) \sim C \exp\Big(-\frac{۱}{۲}||\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}||^۲ - \lambda \sum |\beta_j|^\gamma \Big)$$

نوشته شود.
دقت کنید که در مطالعات بیزی، متداول است که از میانگین پسین[۶] نسبت به مد پسین[۷] (که در اینجا استفاده شده است)، بهره ببرند. این تعدیل (تغییر) از کلیات چیزی را کم نمی‌کند، زیرا تغییرات می‌توانند درون ثابت‌های دلخواه C و λ قرار بگیرند.
قضیه بیز، توزیع‌های پیشین و پسین را به‌صورت

$$f(\boldsymbol{\beta}|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})$$

مرتبط می‌سازد که در آن $\pi(\beta) = f(\boldsymbol{\beta})$ توزیع پیشین $\boldsymbol{\beta}$ است. علاوه بر این، اگر خطای ϵ به‌طور نرمال توزیع شده باشد، آن‌گاه

$$f(\mathbf{Y}|\boldsymbol{\beta}) \propto \exp\Big(-\frac{۱}{۲}||\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}||^۲ \Big).$$

با جبر ساده، نتیجه می‌شود که پیشین‌ها روی هر مولفه $\boldsymbol{\beta}$ به‌صورت

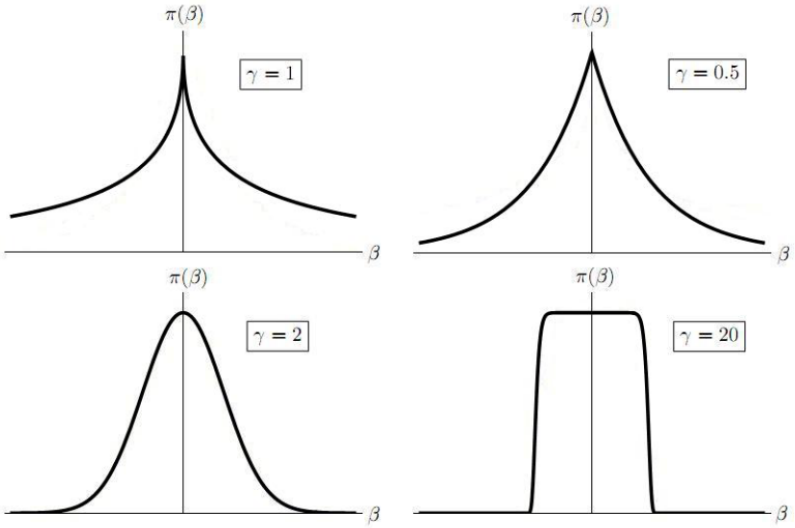
$$\pi(\beta) \propto \exp(-\beta^\gamma)$$

هستند. این پیشین برای تعداد اندکی از مقادیر γ در شکل ۴.۳ رسم شده است و شناخت ارزشمندی درمورد روش کار رگرسیون بریج نشان می‌دهد.

- کمترین توان‌های دوم معمولی دارای پیشین یکنواخت است.
- رگرسیون بریج پیشینی را شامل می‌شود که وزن بیشتری به مقادیر کوچک β نسبت به مقادیر بزرگ آن می‌دهد که این موضوع، باعث انقباض می‌شود.
- برای $\gamma > ۱$ توزیع در ماکسیمم تقریبا صاف[۸] است و بدین معناست که متغیرها تمایل ندارند تا آنجا که ممکن است به $\beta = ۰$ منقبض شوند.

[6] Posterior mean
[7] Posterior mode
[8] Flat

- هنگامی‌که γ افزایش می‌یابد، دم‌های پیشین، منقبض می‌شوند و برای مقادیر بزرگ β، اریبی را درون مدل اضافه می‌کند. تا حدی که برای مقادیر خیلی بزرگ γ، مقادیر کم β اصلا منقبض نمی‌شوند اما مقادیر بزرگ می‌شوند.



شکل ۴.۳: پیشین $\pi(\beta)$ روی β در رگرسیون بریج پیشین $\pi(\beta)$ روی β در رگرسیون بریج، برای مقادیر مختلف γ نشان داده شده است. توجه کنید که این نمودارها همه با یک مقدار یکسان λ به‌دست آمده‌اند ــ در واقعیت رگرسیون بریج به‌ازای هر γ از مقدار متفاوتی از λ برای بهینه‌سازی برازش استفاده می‌کند

طرح متعامد

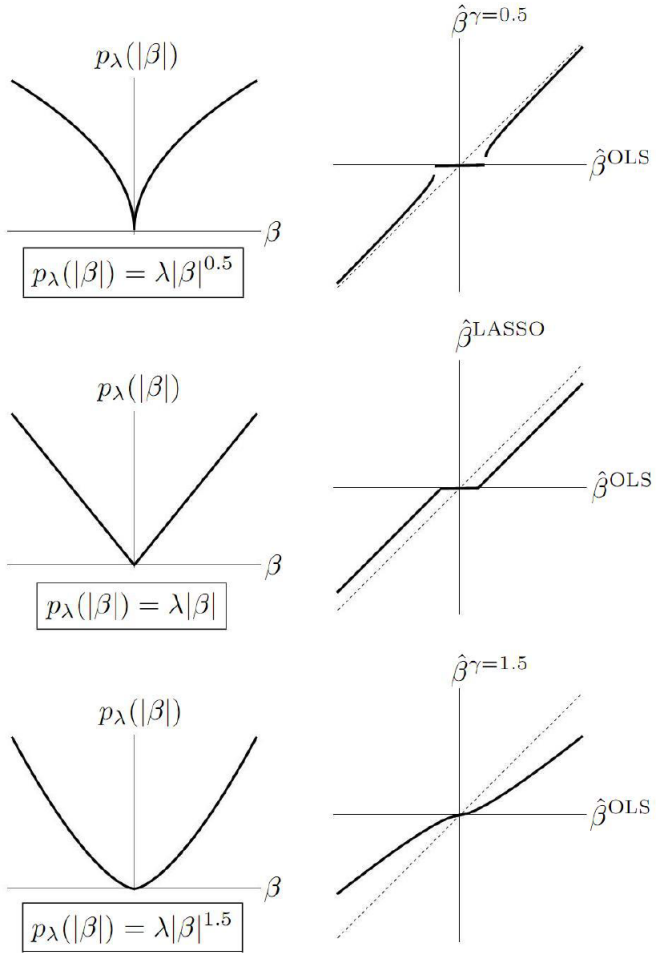
برای به‌دست آوردن شناخت بیشتر درباره رفتار این جریمه‌ها، آن‌ها را در حالت خاص طرح متعامد بررسی میکنیم. در این حالت، جریمه‌ها به‌صورت رابطه (۲.۳) را می‌گیرند.

$$q(\beta_j) = \frac{1}{N}\left(\hat{\beta}_j^{\mathrm{OLS}} - \beta_j\right)^2 + \lambda|\beta_j|^\gamma$$

در ادامه، رفتار این تابع در موارد زیر مورد بررسی قرار می‌گیرد.

- $0 < \gamma < 1$
- $\gamma = 1$
- $1 < \gamma < 2$
- $\gamma = 0$
- $\gamma > 2$

موارد $\gamma = ۱$ و $\gamma = ۲$ جواب‌های تحلیلی را می‌پذیرند. برخی موارد خاص دیگر نیز چنین عمل می‌کنند اما جزئیات دست و پا گیر و ناخوشایند دارند که به‌عنوان یک راه‌حل برای آن‌ها از شبیه‌سازی استفاده می‌شود. یافته‌ها برای جریمه‌های رگرسیون بریج در شکل‌های ۵.۳ و ۶.۳ نمایش داده شده است.

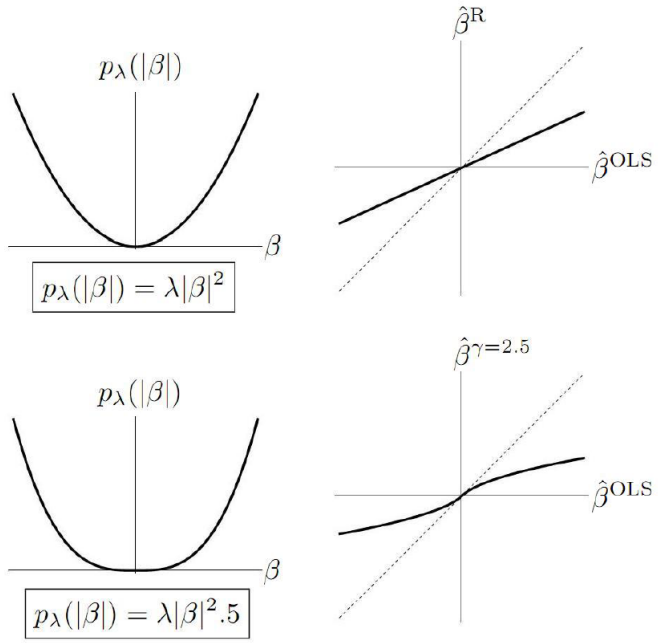


شکل ۵.۳: جریمه‌های بریج و برآوردگرهای به‌دست آمده برای $\gamma = ۰/۵, ۱, ۱/۵$.جریمه‌های بریج و برآوردگرهای به‌دست آمده برای $\gamma = ۰/۵, ۱, ۱/۵$. در هر نمودار، خط نقطه‌چین به‌عنوان مرجع رسم شده است.

بر اساس شکل‌های ۵.۳ و ۶.۳، هر یک از نمودارها مورد بررسی قرار می‌گیرد و نتایج آن‌ها بیان می‌شود. فرض کنید $\Delta = ۲/N$ و برای ساده‌سازی نمادگذاری، $\tilde{\lambda} = \lambda/\Delta$.

$۰ < \gamma < ۱$

متاسفانه، برای هر مقدار γ در این بازه، یافتن پاسخ تحلیلی برای $q(\beta)$ ممکن نیست. برای بررسی رفتار q از شبیه‌سازی استفاده شده است. با توجه به اینکه در نقطه صفر، مشتق‌پذیر نیست، q همیشه دارای کمینه در مبدا است. برای λ به‌حد کافی بزرگ، q همچنین دارای یک نقطه کمینه دیگر است که منجر به رفتار آستانه غیرخطی می‌شود که در شکل ۵.۳ برای



شکل ۶.۳: جریمه‌های بریج و برآوردگرهای به‌دست آمده برای $\gamma = ۲,۲/۵$. جریمه‌های بریج و برآوردگرهای به‌دست آمده برای $\gamma = ۲,۲/۵$. در هر نمودار، خط نقطه‌چین به‌عنوان رسم شده است.

$\gamma = ۰/۵$ نمایش داده شده است. برآوردگر برای β بزرگ ناریب است و انتخاب متغیر را با موفقیت اجرا می‌کند.

$\gamma = ۱$

این، لاسو است. در این حالت

$$q(\beta_j) = \frac{۱}{N}\left(\hat{\beta}_j^{\mathrm{OLS}} - \beta_j\right)^{۲} + \lambda|\beta_j|.$$

کمینه‌سازی چنین تابع مشتق‌ناپذیری مستلزم استفاده از زیرمشتق‌ها است. پاسخ عبارت است از

$$\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}} = \mathrm{sgn}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}})\left(|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}| - \tilde{\lambda}\right)_+$$

که در آن $x_+ = \max(x, ۰)$.

این نوع تابع را در ادبیات آماری آستانه نرم[9] می‌گویند. از شکل ۵.۳ خیلی بدیهی است چگونه لاسو برآورد مقادیر β بزرگتر را اریب می‌کند.

$۱ < \gamma < ۲$

حالت $\gamma = ۱/۵$ در اینجا پاسخ تحلیلی دارد. شکل ۵.۳ به‌وضوح، نشان می‌دهد که در این بازه از γ، مولفه‌های کوچک β با اندازه بزرگی منقبض می‌شوند و مولفه‌های بزرگ β با اندازه

[9] Soft thresholding

کوچکتر منقبض می‌شوند.

$\gamma = ۲$

این رگرسیون، ریج است. در این حالت،

$$q(\beta_j) = \frac{۱}{N}\left(\hat{\beta}_j^{\text{OLS}} - \beta_j\right)^۲ + \lambda|\beta_j|^۲$$

جواب تحلیلی در این حالت، به‌سادگی عبارت است از

$$\hat{\boldsymbol{\beta}} = \frac{۱}{۲\tilde{\lambda}+۱}\hat{\boldsymbol{\beta}}^{\text{OLS}}.$$

به‌عبارت دیگر، رگرسیون ریج، انقباض متناسب را اجرا می‌کند؛ در این جالت، متغیر هرچه بزرگتر باشد، بیشتر منقبض میشود.

$\gamma > ۲$

در این حالت، یکبار دیگر به شبیه‌سازی متوسل شده و نموداری از پاسخ رسم می‌شود.

در این حالت، به‌وضوح مولفه‌های کوچک β حفظ می‌شوند و همان‌طور که از نمودار ۶.۳ به‌ازای $\gamma = ۲/۵$ برمی‌آید، مولفه‌های بزرگتر β منقبض می‌شوند.

خلاصه

رگرسیون بریج، بازه وسیعی از رفتارها هم به‌لحاظ مولفه‌های کوچک و هم بزرگ را در بر می‌گیرد. مولفه‌های کوچک یا کوچکتر از حد آستانه هستند یا خیر. اینکه آیا این اتفاق می‌افتد بستگی به این دارد که آیا جریمه در مبدا تیز است ($\gamma \leq ۱$) یا نیست ($\gamma > ۱$).

انقباض مولفه‌های بیشتر از حد آستانه به تقعر تابع بستگی دارد. تقعر کمتر منجر به انقباض بیشتر مولفه‌های کوچک می‌شود و تقعر بزرگ منجر به انقباض بیشتر مولفه‌های بزرگ می‌شود.

نکات بیشتر درباره رگرسیون ریج

رگرسیون ریج، در واقع رگرسیون بریج با $\gamma = ۲$ است که اولین بار توسط هورل و کنارد (۱۹۷۰) پیشنهاد شده است. آن‌ها تلاش کردند مسئله هم‌خطی در ستون‌های $\mathbf{X}$ را حل کنند. اگر ستون‌های $\mathbf{X}$ هم‌خطی داشته باشند، آن‌گاه دترمینان ماتریس $\mathbf{X}^\top\mathbf{X}$ احتمالا بسیار کوچک است (زیرا، آن توان دوم حجم متوازی‌السطوحی است که لبه‌های آن ستون‌های ماتریس $\mathbf{X}$ هستند). این بدین معنی است که $(\mathbf{X}^\top\mathbf{X})^{-۱}$ احتمالا دارای تعدادی مقدار ویژه بزرگ باشد. این مشکل‌ساز است زیرا این مقادیر ویژه، واریانس‌های مولفه‌های β هستند از دیدگاه مولفه‌های اصلی، مقادیر ویژه کوچک $\mathbf{X}^\top\mathbf{X}$ اشاره بر این دارد که واریانس در راستای برخی مولفه‌های

اصلی کوچک است، یعنی تعیین مدل در آنجا دشوار است. این مشکل مخصوصا در مسائل بعد بالا، که $\mathbf{X}$ دارای تعداد زیادی ستون و تعداد کمی سطر است، با احتمال بالا رخ می دهد.

رگرسیون ریج قصد دارد این مسئله را با کمینه کردن عبارت

$$Q_{۲}(\boldsymbol{\beta}) = \frac{۱}{۲}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^{۲} + \tilde{\lambda}||\boldsymbol{\beta}||^{۲}$$

حل کند.

پاسخ برای این معادله در بین جواب‌های مسائل کمترین توان‌های دوم جریمه‌شده یکتاست. در این حالت، پاسخ می‌تواند به صورت بسته نوشته شود. مشتق گرفتن و برابر صفر قرار دادن منجر به پاسخ

$$\hat{\boldsymbol{\beta}}^{\mathrm{R}} = (\mathbf{X}^{\top}\mathbf{X} + \lambda \boldsymbol{I})^{-۱}\mathbf{X}^{\top}\mathbf{Y}$$

می‌شود. ماتریس هَت نتیجه شده عبارت است از

$$\boldsymbol{H}^{\mathrm{R}} = \mathbf{X}(\mathbf{X}^{\top}\mathbf{X} + \lambda \boldsymbol{I})^{-۱}\mathbf{X}^{\top} \qquad (۹.۳)$$

به‌طور شهودی، رگرسیون ریج به‌آسانی، مقادیر ویژه $\mathbf{X}^{\top}\mathbf{X}$ را افزایش می‌دهد.

می‌توان با در نظر گرفتن تجزیه مقدار ویژه $\hat{\mathbf{Y}}^{\mathrm{R}}$ درک بسیار عمیق‌تری از آن‌چه اتفاق می‌افتد، به‌دست آورد. با توجه به اینکه، تجزیه مقدار ویژه $\mathbf{X}$، برابر $\mathbf{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}$ است، می‌توان نوشت

$$\begin{aligned}
\hat{\mathbf{Y}}^{\mathrm{R}} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\top}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top} + \lambda \boldsymbol{I})^{-۱}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\top}\mathbf{Y} \\
&= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}(\boldsymbol{V}\boldsymbol{D}^{۲}\boldsymbol{V}^{\top} + \lambda \boldsymbol{I})^{-۱}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\top}\mathbf{Y} \\
&= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^{۲} + \lambda \boldsymbol{I})^{-۱}\boldsymbol{D}\boldsymbol{U}^{\top}\mathbf{Y} \\
&= \sum_{j=۱}^{p} \mathbf{u}_j \frac{d_j^{۲}}{d_j^{۲} + \lambda}\mathbf{u}_j^{\top}\mathbf{y}.
\end{aligned}$$

مقایسه این با عبارت مشابه در کمترین توان‌های دوم معمولی رابطه(۱۰.۱)،نتیجه می‌دهد که دقیقا مانند کمترین توان‌های دوم معمولی، رگرسیون ریج، ابتدا $\hat{\mathbf{Y}}$ را بر حسب مولفه‌های اصلی $\mathbf{Y}$ بیان می‌کند. اما پیش از بازگشت به فضای $\hat{\mathbf{Y}}$، مولفه‌ها را در $\frac{d_j^{۲}}{(d_j^{۲}+\lambda)}$ ضرب می‌کند. از این رو، به‌نظر می‌رسد که رگرسیون ریج، تصویر را در راستای همان مولفه‌های اصلی با مقادیر کوچک d_j منقبض می‌کند. این‌ها دقیقا مولفه‌های در راستایی که مشاهدات انتشار نیافته‌اند، هستند (به‌عبارت دیگر، درامتدادی که $\mathbf{X}$ دارای واریانس کمتری است) ـ به‌عبارت دیگر، این‌ها همان مولفه‌هایی هستند که برازش یک خط مستقیم با استفاده از آن‌ها مشکل است و بنابراین، دارای واریانس بالایی خواهند بود.

در نتیجه، رگرسیون ریج تصویر را دقیقا در امتداد آن محورها منقبض می‌کند که در غیر این صورت، واریانس متغیرهای پیشگو بسیار بالا خواهد بود.

همچنین می‌توان این تجزیه مقدار ویژه را برای به‌دست آوردن تعداد پارامترهای موثر (تعریف ۳.۲.۱) در رگرسیون ریج به‌صورت

$$\begin{aligned} d_{\text{eff}^{\text{R}}} &= \operatorname{tr}(\boldsymbol{H}) \\ &= \operatorname{tr}\left(\sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{(d_j^2+\lambda)} \mathbf{u}_j^\top\right) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)} \end{aligned}$$

نوشت. این نتیجه کوچکتر از p است که عدد موثر پارامترها برای کمترین توان‌های دوم معمولی بود. از این رو، در کاهش خوش‌بینی خطای مدل‌ساز در این حالت، با وجود این واقعیت که از تعداد پارامترهای غیرصفر کاسته نشده است، با موفقیت عمل کرده است.

می‌توان نشان داد که رگرسیون منجر به برآورد بهتری (از دیدگاه میانگین توان‌های دوم خطا) از کمترین توان‌های دوم معمولی به‌ازای λ به‌اندازه کافی کوچک منتهی می‌شود (هورل و کنارد، ۱۹۷۰). متاسفانه انتخاب λ خیلی بزرگ می‌تواند به‌طور معنی‌داری، کیفیت مدل را کاهش دهد.

نکات بیشتری درباره رگرسیون لاسو

- برخلاف ریج، لاسو جواب‌های تحلیلی ندارد و برآوردگر لاسو، شکل بسته ندارد. با این وجود، هنوز می‌توان استنباط‌هایی در مورد متغیرهای لاسو انجام داد.

  ممکن است محققی فریفته شود و به‌سادگی بگوید که $d_{\text{eff}}^{\text{LASSO}}$ با تعداد پارامترها در مدل برابر است. اگرچه، این بسیار ساده‌لوحانه است، زیرا آن نتوانسته جستجو برای متغیرهای صحیح را به‌حساب آورد. ژو و همکاران (۲۰۰۷) نشان دادند که تعداد متغیرهای انتخاب شده یک برآوردگر ناریب از مقدار واقعی $d_{\text{eff}}^{\text{LASSO}}$ است.

- لاسو، محدب است و از این رو، می‌تواند با سهولت نسبی محاسبه شود. علاوه بر این، افرون و همکاران (۲۰۰۴)، یک الگوریتم حتی سریع‌تر معرفی کردند که رگرسیون کمترین زاویه[10] (LARS) نامیده می‌شود. این تکنیک به‌خودی خود مفید است و همچنین با یک تغییر جزئی، می‌تواند پاسخ‌های لاسو را پیدا کند.

  روش LARS بسیار شبیه به انتخاب پیشرو است که در آن با یک پاسخ خالی شروع می‌شود و همبسته‌ترین متغیر با پاسخ $\mathbf{Y}$ را اضافه می‌کند ــ برای آسان‌شدن بحث، به آن متغیر X_1 بگویید. با این حال، در روش انتخاب ضریب برای این متغیر، بر خلاف روش انتخاب پیشرو است

[10] Least Angle Regression (LARS)

- انتخاب متغیر پیشرو، ضریب را از بدیهی‌ترین روش، با استفاده از کمترین توان‌های دوم معمولی انتخاب خواهد کرد. این راه منجر به یک برآورد $\hat{\mathbf{Y}}$ خواهد شد با این ویژگی که باقی‌مانده‌های $\mathbf{Y} - \hat{\mathbf{Y}}$ به‌طور کلی با متغیر X_1 ناهمبسته هستند.

  دلیل این رفتار، آن است که کمترین توان‌های معمولی با تصویر کردن بردار $\mathbf{Y}$ روی فضای تولید شده توسط X_1 کار می‌کند ــ از این رو، باقی‌مانده نتیجه شده $\mathbf{Y} - \hat{\mathbf{Y}}$ لزوما عمود بر X_1 است.

- از طرف دیگر، رگرسیون کمترین زاویه رهیافتی دموکراتیک‌تر می‌گیرد.این رگرسیون، برآورد $\hat{\mathbf{Y}}^{\mathrm{LA}}$ را با اضافه کردن هر اندازه از مقدار متغیر X_1 که نیاز باشد، تولید می‌کند تا تضمین کند همبستگی $\hat{\mathbf{Y}}^{\mathrm{LA}}$ با X_1 برابر با همبستگی‌اش با کاراترین متغیر بعدی (بگویید X_2) است. نقطه‌ای که در آن متوقف می‌شود و فرآیند را برای همبسته‌ترین متغیر بعدی (بگویید X_3) تکرار می‌کند.

- ژائو و یو (۲۰۰۶) نشان دادند که تحت شرایط خاص انتخاب متغیر سازگار است (هنگامی که n به‌سمت بی‌نهایت میل می‌کند، مدل درست را با احتمال یک پیدا می‌کند). برخی روش‌ها برای مقابله با موقعیتی که در آن‌ها، این شرط برقرار نباشد، ابداع شده است. یکی از این روش‌ها، لاسو تصادفی‌شده[11] است که در ادامه به آن اشاره می‌شود.

رگرسیون بریج با ماهیت واقعی خود

رگرسیون بریج می‌تواند خودش به‌عنوان یک برآوردگر جریمه‌شده در رقابت با سایر برآوردگرهای جریمه‌شده استفاده شود. منظور این است که با استفاده از روش‌هایی مانند اعتبارسنجی متقابل یا نوع تعمیم‌یافته آن یا معیارهای دیگر، می‌توان هر دو پارامتر λ و $\gamma > 1$ را انتخاب کرد. در واقع، فرانک و فریدمن (۱۹۹۳) برای اولین بار این روش را معرفی کردند. فو (۱۹۹۸) عملکرد بریج را تحلیل کرد و بیان کرد که گاهی اوقات بهتر از لاسو عمل می‌کند (که با توجه به آن‌که لاسو زیرمجموعه از بریج است، منطقی به‌نظر می‌رسد). اما این روش، مملو از سختی است، زیرا غیرخطی بودن برآوردگرهای بریج، اعتبارسنجی متقابل تعمیم یافته را به یک ابزار نامناسب برای انتخاب γ و λ تبدیل می‌کند.

۳.۱.۳ SCAD

همانطور که مطرح شد،روش‌های مدل کلاسیک و جریمه‌های بریج، هیچکدام در هر سه ویژگی تابع جریمه صدق نمی‌کنند. به‌عنوان یک نتیجه، فن و لی (۲۰۰۱) جریمه انحراف قدرمطلق به‌طور هموار کوتاه شده[12] (SCAD) را پیشنهاد دادند، که هر سه شرط را برآورده

[11] Randomized LASSO

[12] Smoothly Clipped Absolute Deviation Penalty (SCAD)

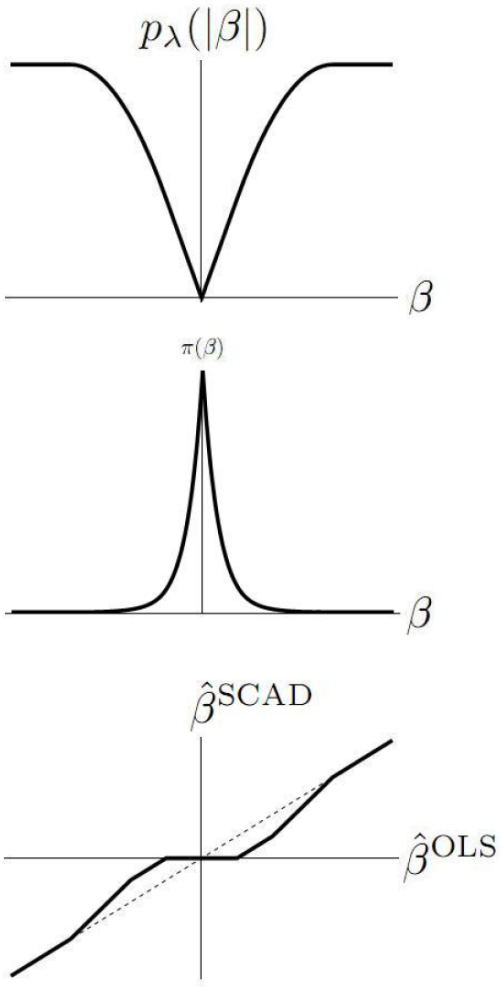
می‌کند. مشتق تابع جریمه SCAD به‌صورت

$$p'_{\lambda}(|\boldsymbol{\beta}|) = \lambda\left\{\mathbb{1}_{|\boldsymbol{\beta}|<\tilde{\lambda}} + \frac{(a\tilde{\lambda}-|\boldsymbol{\beta}|)_+}{(a-1)\tilde{\lambda}}\mathbb{1}_{|\boldsymbol{\beta}|>\tilde{\lambda}}\right\} \qquad (۱۰.۳)$$

تعریف می‌شود. حتی در مورد طرح متعامد، به‌نحوی یافتن جواب سخت است. براوردگر SCAD عبارت است از

$$\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}} = \begin{cases} \mathrm{sgn}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}})\left(|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}| - \tilde{\lambda}\right)_+ & |\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}| \le 2\tilde{\lambda} \\ \frac{(a-1)\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \mathrm{sgn}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}})a\tilde{\lambda}}{(a-2)} & 2\tilde{\lambda} < |\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}| \le a\tilde{\lambda} \\ \hat{\boldsymbol{\beta}}^{\mathrm{OLS}} & |\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}| > a\tilde{\lambda} \end{cases}$$

جریمه SCAD و این نتیجه در شکل ۷.۳ و پیشین بیزی روی β که بر اساس جریمه SCAD به‌دست آمده‌اند، رسم شده‌اند.



شکل ۷.۳: نمودار جریمه، پیشین و برآوردگر SCAD بالا: جریمه p_λ برای SCAD. وسط: پیشین روی β که از جریمه SCAD نتیجه شده است. پایین: برآوردگر SCAD، $\hat{\beta}^{\mathrm{SCAD}}$. خط نقطه‌چین $\hat{\beta}^{\mathrm{SCAD}} = \hat{\beta}^{\mathrm{OLS}}$ است، به‌عنوان مرجع رسم شده است.

شکل ۷.۳، نشان می‌دهد که چگونه SCAD، اریبی لاسو را حل می‌کند. برای مقادیر کوچک β، برآوردگر β را به‌سمت میانگین منقبض می‌کند، اما به برآورد کمترین توان‌های دوم معمولی برمی‌گردد. در چارچوب بیزی، SCAD تیزی ماکسیمم مشاهده شده در لاسو را حفظ می‌کند، اما دم‌ها را ثابت نگه می‌دارد و از این رو اریبی، برای مقادیر بزرگ β کاهش می‌یابد.

تنها مشکل SCAD این است که محدب نیست و از این رو، بهینه‌سازی آن بسیار مشکل است. ژو و لی (۲۰۰۸) تقریب خطی موضعی نزدیک نقطه β_0 زیر را به‌صورت

$$\begin{aligned} p_\lambda(|\boldsymbol{\beta}|) &= p_\lambda(|\beta_0|) + p'_\lambda(|\beta_0|)(|\boldsymbol{\beta}| - |\beta_0|) \\ &= p'_\lambda(|\beta_0|)|\boldsymbol{\beta}| + \text{ثابت} \end{aligned}$$

پیشنهاد کردند. پس می‌توان (SCAD) را با انتخاب نقطه شروع منطقی β_0 و تکرار آن به‌صورت

$$\hat{\boldsymbol{\beta}}^{k+1} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{N} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}|| + \sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^k|)|\beta_j| \right)$$

حل کرد. به‌طور قابل توجهی، برآوردگر اولیه فراهم شده به‌طور معقولانه‌ای خوب است (برای مثال، با استفاده از لاسو به‌دست آمده است)، معلوم است که یک گام از الگوریتم، به‌خوبی کل الگوریتم تکراری است. علاوه بر این، تحت شرایط نظم معین، می‌توان نشان داد که در این برآوردی خاصیت پیش‌گویی صدق می‌کند. یک رویه انتخاب متغیر دارای خاصیت پیش‌گویی است اگر بتواند زیرمجموعه درست و واقعی متغیرها را شناسایی کند. فن و لی (۲۰۰۱) نیز یک تقریب توان دوم موضعی برای SCAD پیشنهاد دادند.

یک پیچیدگی اضافی از این حقیقت ناشی می‌شود که در ساختار تابع جریمه SCAD، دو ثابت نامشخص λ و a وجود دارند. هر دو می‌توانند با استفاده از روش‌های اعتبارسنجی متقابل برآورد شوند اما از نظر محاسباتی، بسیار پیچیده است. در عوض، فن و لی (۲۰۰۱) استدلال بیزی زیر را برای ثابت کردن a پیشنهاد دادند:

- قرار دهید $\lambda = \sqrt{2\log(p)}$. به این مقدار آستانه کلی [13] می‌گویند که توسط دونوهو و جانستون (۱۹۹۴) پیشنهاد شده بود.
- فرض کنید $\boldsymbol{\beta}$ دارای پیشین نرمال با میانگین صفر و واریانس $a\lambda$ باشد.
- مخاطره بیزی را برای هر مقدار a با انتگرال گیری عددی محاسبه کنید. به‌سادگی، مخاطره بیزی عبارت است از
$$\mathrm{E}\Big\{ L(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \Big\}$$
- شبیه‌سازی را برای تعدادی از مقادیر p انجام دهید.

در همه حالت‌ها، کمترین مخاطره در $a \approx ۳/۷$ اتفاق می‌افتد و در شبیه‌سازی، این مقدار شبیه به مقادیر a به‌دست آمده از اعتبارسنجی متقابل عمل می‌کند.

[13] Universal thresholding

۴.۱.۳ انتخاب پارامتر تنظیم‌کننده λ

همه روش‌هایی که تاکنون مطرح شده است، یک پارامتر تنظیم‌کننده λ دارند. که شدت جریمه p_λ را تنظیم می‌کند. λ نیز مشخص می‌کند در چه نقطه‌ای از طیف اریبی–واریانس قرار دارد. مقادیر کوچک λ، جریمه را کاهش می‌دهد و از این رو، مدل‌های پیچیده‌تر را مجاز می‌داند ــ این اریبی را کاهش میدهد اما واریانس را افزایش می‌دهد.

فرض کنید $\hat{\mathcal{M}}^\lambda$ نشان دهنده مدل به‌دست آمده از یک روش مشخص با پارامتر تنظیم‌کننده λ باشد و فرض کنید Λ مجموعه همه پارامترهای ممکن برای λ باشد. وظیفه یک محقق، انتخاب یک مدل (بهترین) از بین مجموعه همه مدل‌های ممکن

$$\{\hat{\mathcal{M}}^\lambda; \lambda \in \Lambda\}$$

است. در این بخش، روش‌هایی برای انجام آن، مطرح می‌شود.

اعتبارسنجی متقابل و اعتبارسنجی متقابل تعمیم‌یافته

رایج‌ترین رهیافت برای انتخاب λ، اعتبارسنجی متقابل و اعتبارسنجی متقابل تعمیم‌یافته است. به‌طور موثری مجموعه Λ با تعدادی از مقادیر گسسته از قبل تعیین شده برای λ پر شده است. برای هر یک از این مقادیر، امتیاز اعتبارسنجی متقابل به‌کار گرفته شده است که به‌صورت زیر تعریف شده است.

تعریف ۲.۱.۳. [امتیاز اعتبارسنجی متقابل] مجموعه مدل‌ساز $\mathcal{T}$ به K قسمت یا دسته مساوی تقسیم می‌شود. برای نمایش مقدار برازش شده y وقتی مدل با پارامتر تنظیم‌کننده λ، به $\mathcal{T}$ منهای قسمت شامل داده iام برازش شده، از نماد $\hat{Y}^{-k(i),\lambda}$ استفاده می‌شود. سپس، امتیاز اعتبارسنجی متقابل عبارت است از

$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\big(Y_i, \hat{Y}^{-k(i),\lambda}\big)$$

سرانجام، مقدار $\lambda \in \Lambda$ که کمترین امتیاز اعتبارسنجی متقابل را تولید کند، انتخاب می‌شود.

در برخی موارد، محاسبه کردن امتیاز اعتبارسنجی متقابل تعمیم‌یافته ممکن است، که معمولا به محاسبه کمتری نیاز دارد. برای مثال، برای حالت خاص رگرسیون ریج، گلوب و همکارن (۱۹۷۹) نشان دادند که کمینه کردن امتیاز اعتبارسنجی متقابل تعمیم‌یافته دقیقا به‌خوبی کمینه کردن امتیاز اعتبارسنجی متقابل کار می‌کند.

$$\mathrm{GCV}(\hat{f}) = N\frac{||\mathbf{Y} - \boldsymbol{H}^{\mathrm{R}}\mathbf{Y}||^2}{(1 - \mathrm{tr}(\boldsymbol{H}^{\mathrm{R}}))^2}$$

که در آن $\boldsymbol{H}^{\mathrm{R}}$ ماتریس هَت برای برآورد ریج تعریف شده در معادله (۹.۳) است. این معبار بسیار آسان‌تر از امتیاز اعتبارسنجی متقابل محاسبه می‌شود.

انتخاب پایدار

اعتبارسنجی متقابل، یک مدل را از بین اعضای مجموعه

$$\{\hat{\mathcal{M}}^{\lambda}; \lambda \in \Lambda\}$$

انتخاب می‌کند، با این حال، این راه ممکن است بهترین رهیافت نباشد، زیرا ممکن است که بهترین مدل، حتی در این مجموعه نباشد.

انتخاب پایدار، تاحدودی یادآور روش بوت‌استرپ است. این روش مکررا داده را آشفته می‌کند و متغیرهایی که در سهم بزرگی از مدل‌های نتیجه شده حضور دارند را جستجو می‌کند. این‌ها متغیرهایی هستند که برای شکل دادن بخشی از مدل نهایی، $\hat{\mathcal{M}}^{\text{stable}}$ انتخاب شده‌اند.

تعریف ۳.۱.۳ (احتمال‌های انتخاب). فرض کنید I یک نمونه تصادفی از داده‌ها به اندازه $\lfloor n/۲ \rfloor$ باشد (که $\lfloor x \rfloor$ بزرگترین عدد صحیح کوچکتر یا مساوی با x است) که بدون جایگزینی انتخاب شده‌اند. فرض کنید $\hat{\mathcal{M}}^{\lambda}(I)$ نتیجه استفاده از یک روش با پارامتر تنظیم‌کننده λ روی داده‌های I باشد.

اکنون هر متغیر k در مسئله را مدنظر قرار دهید. احتمال اینکه این متغیر انتخاب شود، به‌صورت

$$\hat{\prod}_{k}^{\lambda} = P\Big(k \subseteq \hat{\mathcal{M}}^{\lambda}(I)\Big)$$

نشان داده می‌شود. احتمال P روی همه زیرنمونه‌های ممکن I گرفته شده است.

اکنون، مدل پایدار به‌صورت زیر تعریف می‌شود.

تعریف ۴.۱.۳ (مدل پایدار). مدل پایدار $\hat{\mathcal{M}}^{\text{stable}}$ به‌صورت

$$\hat{\mathcal{M}}^{\text{stable}} = \Big\{k : \max_{\lambda \in \Lambda} \hat{\prod}_{k}^{\lambda} \geq \pi\Big\}$$

انتخاب می‌شود که در آن π ($۰ < \pi < ۱$) یک مقدار برش است.

به‌عبارت دیگر، متغیرهای با احتمال بالای انتخاب حفظ می‌شود و آن‌هایی که با احتمال پایین، انتخاب می‌شوند، از مدل حذف می‌شوند.

بحث چگونگی انتخاب π و Λ باقی می‌ماند. با این حال، قبل از انجام دادن، شایان ذکر است که مینشاوسن و بولمن (۲۰۱۰) به‌طور تجربی نشان دادند که نتایج برای انتخاب‌های معقول در بازه برش $\pi \in (۰/۶, ۰/۹)$، کمی متفاوت است.

در انتخاب π، تلاش می‌شود تا نرخ خطای هر خانواده [۱۴] V کاهش یابد.

تعریف ۵.۱.۳ (نرخ خطای خانواده). نرخ خطای خانواده، V، به‌عنوان تعداد متغیرهای مزاحم انتخاب شده در $\hat{\mathcal{M}}^{\text{stable}}$ تعریف شده است. به‌عبارت دیگر، تعداد متغیرهایی است که در مدل موردنظر انتخاب شده‌اند ولی بر روی Y تاثیر نمی‌گذارند.

[14] Per-family error rate

تعریف ۶.۱.۳ (بازه انتخاب Λ)**.** بازه انتخاب Λ، q_Λ، تعداد کل متغیرهایی است که روش پیشنهادی قادر به انتخاب از زیرنمونه‌هایی است به‌طوری که مدل، هر مقدار در مجموعه Λ را اختیار کند.

$$q_\Lambda = \mathrm{E}\Big(\Big|\cup_{\lambda\in\Lambda}\hat{\mathcal{M}}^\lambda(I)\Big|\Big)$$

.

بنا به روش استفاده شده برای برازش مدل، نسبت دادن q_Λ به Λ باید نسبتا ساده باشد. برای مثال، برای لاسو، λ کوچکتر به‌معنی متغیرهای کمتر است، بنابراین اگر $\lambda_{\min}$، کوچکترین عضو Λ باشد، q_Λ، در واقع، تعداد متغیرها در مدل $\hat{\mathcal{M}}^\lambda_{\min}$ است.

قضیه ۱.۱.۳. تحت پذیره‌های مشخص (مینشاوسن و بولمن، ۲۰۱۰)، نرخ خطا در هر خانواده به‌صورت

$$\mathrm{E}(V) \le \frac{۱}{۲\pi - ۱}\frac{q_\Lambda^۲}{p} \qquad (۱۱.۳)$$

کران‌دار است که در آن، p تعداد کل متغیرهای حاضر در مدل است.

با کمال تعجب، قضیه پیش‌بینی می‌کند که با افزایش آستانه π و کاهش تعداد کل متغیرها در مدل‌هایی که ممکن است q_Λ متغیر را انتخاب کند، احتمال انتخاب یک متغیر مزاحم کاهش پیدا می‌کند.

تاکتیک در این‌جا این است که یکی از q_Λ و π تعیین شود و سپس از رابطه (۱۱.۳) برای تعیین دیگری به‌منظور به‌دست آوردن مقدار مورد انتظار نرخ خطا در هر خانواده، استفاده شود. یک رهیافت متداول، این است که $\pi = ۰/۹$ قرار داده شود و سپس متناظر با آن، q_Λ (و بنابراین، Λ) تعیین شود.

همچنین دقت کنید که در بعضی موارد، روش خاص استفاده شده برای برازش، هزینه محاسباتی بالایی دارد و از این جهت، استفاده از روش انتخاب پایدار برای تعداد زیاد مقادیر λ غیرعملی می‌شود. در چنین حالتی، انتخاب یک مقدار λ و قرار دادن $\Lambda = \lambda$ امکان‌پذیر است. مینشاوسن و بولمن (۲۰۱۰) به‌طور تجربی نشان دادند این روش بسیار موفق می‌باشد به شرطی که یک مقدار λ طوری انتخاب شده باشد که مقداری بیش برازشی اتفاق افتد (یعنی به‌طوری که مدل $\hat{\mathcal{M}}^\lambda$ خیلی بزرگ باشد).

۲.۳ پیشرفت‌ها

در مطالعاتی که تاکنون انجام شده‌اند، پیشرفت‌های بسیاری در روش‌هایی که تاکنون مطرح شدند، برای رویارویی با حالت غیراستاندارد، دیده می‌شود. در ادامه، دو روش مطرح می‌شود:

الاستیک‌نت (ژو و هیستی، ۲۰۰۵)، غربال‌گری مستقل مطمئن[15] (SIS) (فن و ال‌وی، ۲۰۰۸). برخی دیگر، در ارتباط با این موضوع، لاسو سازوار (ژو، ۲۰۰۶) و انتخاب‌گر دنزینگ (کاندس و تائو، ۲۰۰۷) هستند.

۱.۲.۳ رویارویی با متغیرهای همبسته ـ الاستیک نت

ژو و هیستی (۲۰۰۵) برآوردگر الاستیک‌نت را به‌عنوان راه‌حلی برای دو مسئله زیر پیشنهاد کردند.

۱. لاسو، هیچ‌گاه قادر به ساخت یک مدل با بیش از n متغیر نیست به‌طوری که n، تعداد داده‌های در دسترس است (افرون و همکاران، ۲۰۰۴). اگر بعد مسئله، p، بسیار بزرگتر از n باشد (برای مثال، در تحلیل‌های ژنتیک که در آن هزاران ژن بر اساس کمتر از ده آزمایش میکروآرایه، به‌دست آمده‌اند)، این می‌تواند مشکل ساز باشد.

   ریج قادر به انتخاب مدل‌های بزرگتر است اما انتخاب متغیر را انجام نمی‌دهد (یعنی تُنُک نیست).

۲. اگر یک مدل، شامل تعدای از متغیرهای همبسته باشد، لاسو با احتمال بالایی، یکی از این متغیرهای همبسته را برمی‌گزینند. در بسیاری از مسئله‌های کاربردی، این مطلوب است. به‌عنوان مثال، در مدل‌های با بعد بسیار بالا، احتمال دارد که هر متغیر معنی دار با تعدادی متغیر مزاحم، همبستگی داشته باشد. در آن حالت، بهتر است که فقط معنی‌دارترین متغیر انتخاب شود.

   با این وجود، این موضوع می‌تواند در مسئله‌های کاربردی مشکل‌ساز باشد که به این دلیل است که متغیرهای همبسته از یک گروه هستند. به‌عنوان مثال، در مطالعات ژنتیک ژن‌هایی که یک مسیر متابولیک خاصی را تنظیم می‌کنند، به‌شدت همبسته هستند و اگر مسیر متابولیک، پاسخ را تحت تاثیر قرار دهد، باید همه آن‌ها در مدل وجود داشته باشند. ریج، این رفتار را از خود نشان نمی‌دهند. درحقیقت، تمایل به انتخاب متغیرهای همبسته با یکدیگر در یک مدل دارند. هر چند ریج انتخاب متغیر را انجام نمی‌دهد.

هدف، یافتن جریمه‌ای است که هم تُنُک است و هم بتواند متغیرهای همبسته با یکدیگر را در یک گروه قرار دهد.

در ادامه، به‌اختصار رفتار کمترین توان‌های دوم جریمه‌شده نسبت به متغیرهای همبسته بررسی می‌شود.

[15] Sure Independence Screening (SIS)

متغیرهای همبسته

ژو و هیستی (۲۰۰۵) و تیبشیرانی (۱۹۹۶) از جواب‌های تحلیلی برای بررسی نحوه رفتار مسیرهای جواب لاسو و ریج در رویارویی با متغیرهای همبسته استفاده کردند. تحلیل‌های آن‌ها محدود به حالت دو متغیره می‌شود. (افرون و همکاران، ۲۰۰۴) با معرفی الگوریتم LASR این بینش ارائه کردند که چرا لاسو فقط یک متغیر را از بین یک گروه از متغیرهای همبسته انتخاب می‌کند. در واقع، وقتی بهترین متغیر در گروه به مسیر LARS اضافه می‌شود، همبسته بودن باقی‌مانده‌ها با متغیرهای باقی‌مانده در گروه همبسته بعید است.

در ادامه، یک توضیح هندسی از این رفتار بر حسب تابع جریمه لاسو ارائه می‌شود. پاسخ مسئله کمترین توان‌های دوم جریمه‌شده را به‌صورت

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \sum_{i=1}^{p} \lambda p(|\beta_i|)$$

درنظر بگیرید.براساس نظریه بهینه‌سازی محدب (بوید و ون‌دن‌برگ، ۲۰۰۴) این مسئله برای برخی مقادیر t با مسئله بهینه‌سازی زیر یکسان است

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 \quad \text{such that} \quad \sum_{i=1}^{p} p(|\beta_i|) \leq t.$$

در اولین فرمول‌بندی، λ می‌تواند به‌عنوان یک ضریب لاگرانژ درنظر گرفته شود.

این مسئله بهینه‌سازی تلاش می‌کند تا حد امکان به برآورد کمترین توان‌های دوم معمولی $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ نزدیک شود درحالی که درون محدودیت باقی می‌ماند. بنابراین، پاسخ در اشتراک ترازهای $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ و تراز متناظر با $\sum_{i=1}^{p} p(|\beta_i|) = t$ قرار خواهد گرفت. صورت این ترازها را در نظر بگیرید.

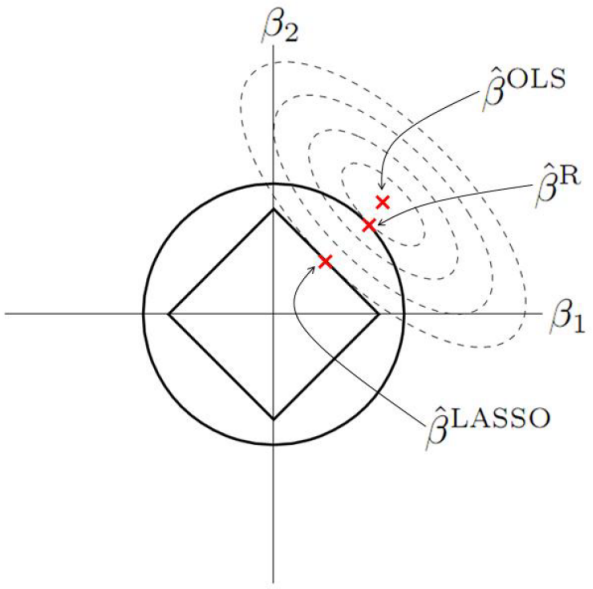
- بدیهی است که ترازهای ثابت $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ بیضی هستند و کمینه (یعنی، مرکز بیضی) به‌وضوح، برآورد کمترین توان‌های معمولی از $\boldsymbol{\beta}$ است.
- ترازهای ثابت $\sum_{i=1}^{p} p(|\beta_i|) \leq t$ به نوع جریمه استفاده شده، بستگی دارد. برای جریمه ریج $(p(|\boldsymbol{\beta}|) = |\boldsymbol{\beta}|^2)$، ترازها دایره‌هایی هستند که مرکز آن‌ها، مبدا محور مختصات است و برای جریمه لاسو، $(p(|\boldsymbol{\beta}|) = |\boldsymbol{\beta}|)$ ترازها الماس[۱۶]هایی هستند که مرکز آن‌ها مرکز محور مختصات هستند.

این مفاهیم در شکل ۸.۳ نمایش داده شده‌است.

اکنون، هندسه ترازهای $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ با جزئیات بیشتری مطرح می‌شود. ابتدا توجه کنید

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 = \mathbf{Y}^\top\mathbf{Y} - 2\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

[16] Diamond



شکل ۸.۳: کمترین توان‌های دوم جریمه شده به‌عنوان یک مسئله بهینه‌سازی محدود شده. کمترین توان‌های دوم جریمه شده به‌عنوان یک مسئله بهینه‌سازی محدود شده. خطوط نقطه‌چین ترازهای $||\mathbf{Y}-\mathbf{X}\beta||^2$ هستند (در برآورد کمترین توان‌های دوم معمولی، مرکزی شده‌اند) و خطوط ممتد، ترازهای توابع جریمه لاسو و ریج در یک ثابت معین t هستند. پاسخ‌ها در تقاطع ترازها اتفاق می‌افتد.

دقت کنید که $\mathbf{X}^\top\mathbf{X}$، ماتریس واریانس‌ـ‌کوواریانس $\mathbf{X}$ است. هندسه بیضی‌ها نتیجه می‌دهد که شبه‌ـ‌محور بلند (قطر بزرگ) ترازهای بیضوی به‌طور معکوس با مقادیر ویژه این ماتریس ارتباط دارند.

در راستای درک بهتر، به هندسه بیضی‌ها اشاره کوتاهی می‌شود. شکل کلی یک بیضی، به‌صورت

$$AX^2 + 2BXY + CY^2 + 2DX + 2FY + M = 0$$

نوشته شوند که به شکل ماتریسی (که می‌تواند به بعدهای بالاتر تعمیم داده شود)، به‌صورت

$$\mathbf{X}^\top \boldsymbol{A}\mathbf{X} + 2\boldsymbol{J}^\top\mathbf{X} + M = 0$$

نوشته شوند به‌طوری که

$$\boldsymbol{A} = \begin{pmatrix} A & B \\ B & C \end{pmatrix} \qquad \boldsymbol{J} = \begin{pmatrix} D \\ F \end{pmatrix}.$$

در اولین گام، ماتریس $\boldsymbol{A}$ باید قطری‌سازی شود و به‌صورت

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}^2\boldsymbol{P}^\top$$

نوشته شود که $\boldsymbol{D}$ یک ماتریس قطری است و $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$.

سپس، یک سیستم مختصات جدید $\tilde{\mathbf{X}}$ تعریف می‌شود که $\tilde{\mathbf{X}} = \boldsymbol{P}^\top\mathbf{X}$ (که معادل با دوران محورهای مختصات است تا در راستای مختصات بیضی قرار بگیرید، یعنی، با مقادیر ویژه

ماتریس $\boldsymbol{A}$). می‌توان بیضی را به‌صورت

$$\tilde{\mathbf{X}}^\top \boldsymbol{D}^2 \tilde{\mathbf{X}} + 2\boldsymbol{J}^\top \boldsymbol{P}\tilde{\mathbf{X}} + M = 0$$

بازنویسی کرد. برای کامل کردن مربع، نتیجه می‌شود که

$$\left(\boldsymbol{D}\tilde{\mathbf{X}} + \boldsymbol{D}^{-1}\boldsymbol{J}^\top \boldsymbol{P}\right)^2 - \boldsymbol{P}^\top \boldsymbol{J}\boldsymbol{D}^{-2}\boldsymbol{J}^\top \boldsymbol{P} + M = 0.$$

سرانجام، یک سیستم مختصات جدید $\hat{\mathbf{X}}$ تعریف می‌شود که $\tilde{\mathbf{X}} = \hat{\mathbf{X}} - \boldsymbol{D}^{-2}\boldsymbol{J}^\top \boldsymbol{P}$. آن‌گاه، بیضی به‌صورت

$$\left(\boldsymbol{D}\hat{\mathbf{X}}\right)^2 = \text{ثابت}$$

که به‌سادگی نتیجه می‌دهد که شبه‌ـمحورهای بیضی، در واقع، برابر با معکوس مقادیر روی قطر ماتریس $\boldsymbol{D}$ است که مقادیر ویژه ماتریس $\boldsymbol{A}$ می‌باشد.

اکنون، حالت دو متغیره در نظر گرفته می‌شود. فرض کنید شرایط نرمال‌سازی در تعریف (۱.۲.۱) برقرار هستند، ماتریس واریانس‌ـ کوواریانس به‌صورت

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

است. مقادیر ویژه این ماتریس عبارتند از

$$1 - \rho \qquad \text{و} \qquad 1 + \rho$$

و نیمه محورهای (قطرهای) بیضی، با

$$\frac{1}{1-\rho} \qquad \text{و} \qquad \frac{1}{1+\rho}$$

متناسب هستند.

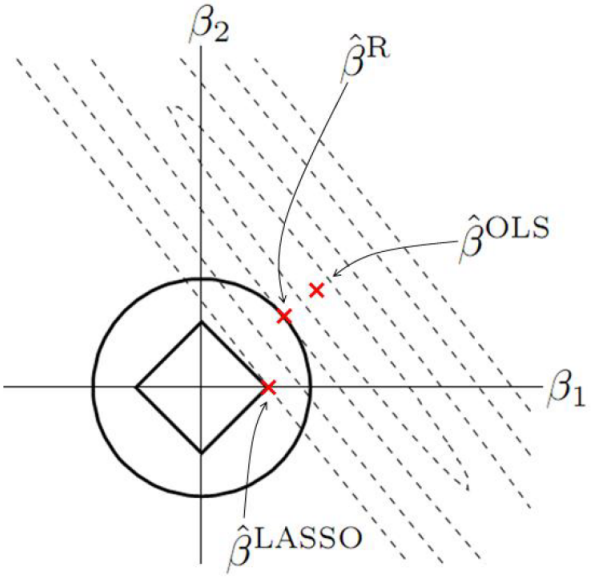
این بدین معنی است که متغیرها بیشتر همبسته می‌شوند (یعنی، $\rho \to$)، اگر بیضی بیشتر و بیشتر کشیده‌تر شود.

اکنون، درک رفتار لاسو و ریج در رویارویی با متغیرهای همبسته و در نتیجه بیضی‌های خیلی کشیده، آسان‌تر است.

- ریج (یا در واقع، هر تابع به‌شدت محدب) برای برخورد با بیضی متورم می‌شود ـ بنابراین، پاسخ دورتر از محورها اتفاق می‌افتد و هر دو متغیر را شامل می‌شود.
- لاسو (یا هر جریمه مقعر دیگر) برای برخورد با بیضی متورم نمی‌شود.

در حالت متغیرهای ناهمبسته، این تفاوتی ایجاد نمی‌کند، زیرا خود بیضی متورم خواهد شد و لاسو، به‌احتمال زیاد هردو متغیر را انتخاب می‌کند (برای مثال، در شکل ۸.۳).

با این حال، اگر متغیرها همبسته باشند، بیضی نیز خیلی کشیده خواهد بود و بستگی به شیب بیضی، لاسو یکی از متغیرهای همبسته را انتخاب خواهد کرد.



شکل ۹.۳: نمودار لاسو و ریج برای دو متغیر با همبستگی بالا لاسو و ریج برای دو متغیر با همبستگی بالا ($\rho = -۰/۹۹$). خطوط نقطه‌چین ترازهای $||\mathbf{Y} - \mathbf{X}\beta||^۲$ هستند. ریج به‌سمت بیرون متورم می‌شود و جوابی شامل هر دو متغیر را انتخاب می‌کند. لاسو، متورم نمی‌شود و فقط یکی از دو متغیر را انتخاب می‌کند (در این مورد $\beta_۱$).

این پدیده، در حالت همبستگی بالا در شکل ۹.۳ نمایش داده شده است.

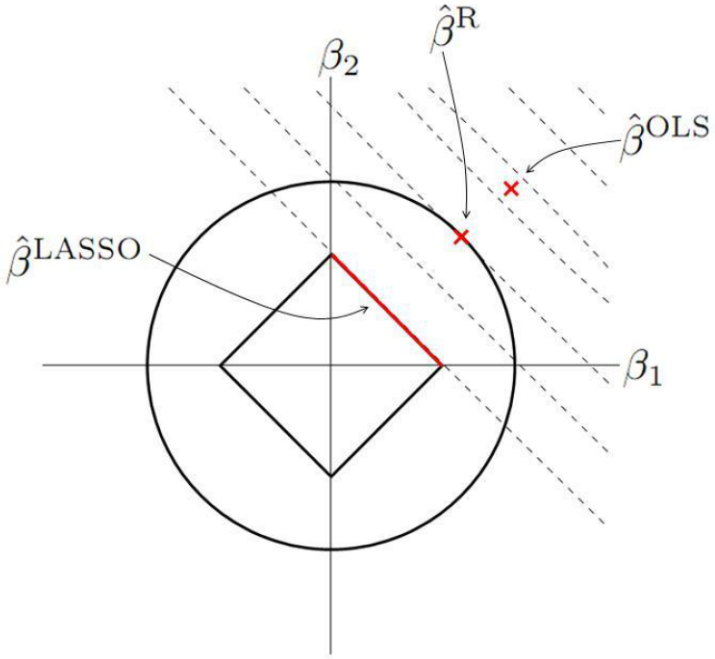
توجه کنید که شکل ۹.۳ کمی گمراه‌کننده است زیرا ماتریس طرح $\mathbf{X}$ استفاده شده برای تولید آن در شرط‌های نرمال‌سازی تعریف ۱.۲.۱ صدق نمی‌کند.

برای درک اینکه چرا ضروری است که ماتریس $\mathbf{X}$ در شرط نرمال‌سازی صدق کند، یادآوری می‌شود که شرایط نرمال‌سازی در حالت دو متغیره مستلزم آن است که

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} ۱ & \rho \\ \rho & ۱ \end{pmatrix}.$$

هم‌چنین، بردارهای ویژه آن ماتریس در راستای محور بیضی قرار می‌گیرد. بردارهای ویژه این ماتریس خاص، $(۱,۱)^\top$ و $(۱,-۱)^\top$ هستند و بنابراین بیضی در زاویه ۴۵ درجه با محورها قرار می‌گیرد. این بدین‌معنی است که بیضی، می‌تواند دقیقا با تابع جریمه لاسو، موازی شود. شکل ۱۰.۳ را برای نموداری که می‌تواند نتیجه شود را ببینید.

شکل ۱۰.۳، آن را در این حالت بدیهی می‌سازد که تعدادی پاسخ برای لاسو ممکن است وجود داشته باشد و بیشتر آن‌ها، هر دو پارامتر $\beta_۱$ و $\beta_۲$ را شامل می‌شوند. این حالت خاص در دو بعد را می‌توان یک استثنا دانست که لاسو، یکی از گروه متغیرهای همبسته برمی‌دارد. برای نشان دادن نقطه تقاطع، بهتر است از ماتریس نرمال‌ـنشده $\mathbf{X}$ استفاده شود.



شکل ۱۰.۳: شکل ۹.۳ در حالت نرمال سازی شده ماتریس $\mathbf{X}$ شکل ۹.۳ را در موقعیتی نشان می‌دهد که ماتریس $\mathbf{X}$ در شرایط نرمال‌سازی در تعریف ۱.۲.۱ صدق می‌کند. بدیهی است، چندین پاسخ ممکن برای لاسو وجود دارد که بیشتر آن‌ها شامل هر دو متغیر هستند.

بنابراین، یک شرط لازم برای انتخاب گروه متغیرهای همبسته این است که تابع جریمه باید اکیدا محدب[17] باشد.

پاسخ

در نگاه اول، به‌نظر می‌رسد که این مسئله می‌تواند با استفاده از رگرسیون بریج به‌ازای مقدار $1 < \gamma < 2$ حل شود ــ این توابع اکیدا محدب هستند. با این حال، همان‌طور که در جدول ۱.۳ مشاهده شد، رگرسیون بریج، فقط برای $\gamma \leq 1$ تُنُک است.

ژو و هیستی (۲۰۰۵) استفاده از یک ترکیب خطی از لاسو و ریج به‌صورت

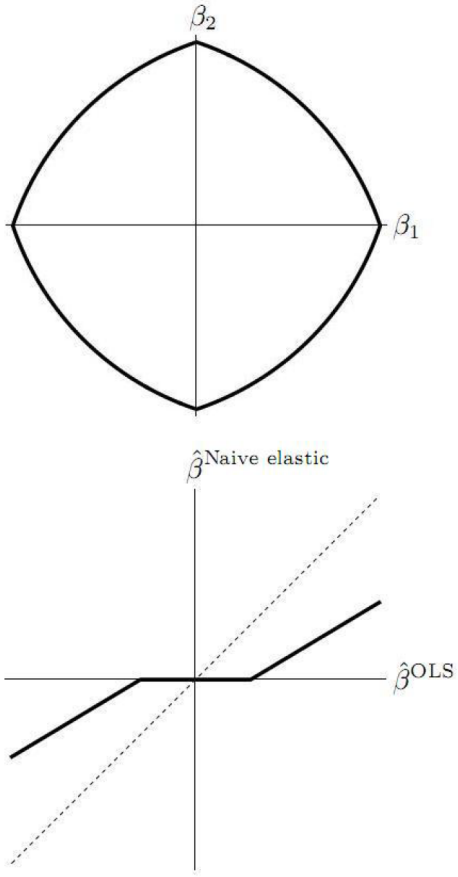
$$p_\lambda(|\boldsymbol{\beta}|) = \lambda_1|\boldsymbol{\beta}| + \lambda_2|\boldsymbol{\beta}|^2 \qquad (۱۲.۳)$$

را پیشنهاد کردند که آن را الاستیک‌نت نامیدند.

پاسخ به‌راحتی از نتایج لاسو و ریج محاسبه می‌شود. شکل تابع جریمه و برآوردگر الاستیک‌نت در شکل ۱۱.۳ نشان داده شده است. توجه کنید که تابع جریمه هنوز دارای گوشه‌ها در مبدا است، اما اکیدا محدب است. این همان چیزی است که هر دو ویژگی مورد نظر را نتیجه می‌دهد.

شکل ۱۱.۳ مشخص می‌کند که الاستیک‌نت معادل با انجام انقباض ریج به‌دنبال آستانه

[17] Strict convexity



شکل ۱۱.۳: ترازهای برآوردگر الاستیک‌نت بالا: ترازهای جریمه الاستیک‌نت ساده (در حالت دو متغیره). پایین: اثر الاستیک‌نت ساده در حالت طرح متعامد.

لاسو است. متاسفانه این انقباض مضاعف[18] نتایج زیر-بهینه تولید می‌کند، زیرا محقق را به خیلی دورتر از موازنه اریبی، واریانس سوق می‌دهد.

به‌عنوان یک نتیجه، ژو و هیستی (۲۰۰۵) روش مطرح شده را الاستیک‌نت ساده[19] نامیدند و یک برآوردگر تصحیح‌شده به‌صورت

$$\hat{\boldsymbol{\beta}}^{\text{Elastic}} = (1 + \lambda_2)\hat{\boldsymbol{\beta}}^{\text{Naiveelastic}}$$

پیشنهاد کردند. عامل $1 + \lambda_2$، دارای اثر انجام ندادن برخی از انقباض‌ها است و در برخی راه‌حل‌ها به‌شدت کاربردی است (ژو و هیستی، ۲۰۰۵).

۲.۲.۳ رویارویی با مسائل بعد بسیار بالا ـ غربال‌گری مستقل مطمئن

در این بخش، مسائل بعد بسیار بالا، یعنی مسائلی که در آن بردارهای ورودی $\mathbf{X}$ به‌حدی بزرگ هستند که حتی روش‌هایی که تاکنون بحث شده‌اند قادر به برازش مدل نیستند، مطرح

[18] Double-shrinkage
[19] Naive elastic net

می‌شود. این می‌تواند به چند دلیل اتفاق بیافتد. دو مورد از بدیهی‌ترین دلایل عبارتند از

- الگوریتم‌های بهینه‌سازی محدب معمولا در فضاهای بعد بسیار بالا به‌طول می‌انجامند.
- وقتی تعداد متغیرها بسیار زیاد افزایش می‌یابد، بسیار محتمل می‌شود که تعدادی از متغیرها با یکدیگر همبسته خواهند بود (فن و ال‌وی، ۲۰۰۸). دانستن اینکه کدام یک از متغیرها واقعا بر Y اثر می‌گذارند و کدام متغیرها فقط به‌طور فریبنده با Y همبسته هستند، بسیار سخت است. این موضوع، مدل با تشخیص‌پذیری[20] کمتر می‌سازد.

برای یک بحث کامل‌تر از مسائل ابعاد بالا، فن و ال‌وی (۲۰۰۰) و فن و لی (۲۰۰۶) را ببینید.

غربال‌گری مستقل مطمئن (فن و ال‌وی، ۲۰۰۸) یک رهیافت ساده و سریع برای هَرس[21] برخی از متغیرهای در مسائل بعد بسیار بالا است. سپس مدل نتیجه شده، به‌حدی کوچک است که کاربرد روش‌های مطرح‌شده را اجازه دهد.

SIS پایه‌ای

روش استفاده‌شده در غربالگری مستقل مطمئن به‌طور فریبنده‌ای ساده است. به‌سادگی، متغیرهایی که به‌شدت با پاسخ Y همبسته‌اند را انتخاب می‌کند. به‌طور رسمی‌تر، فرض کنید

$$\omega = \mathbf{X}^\top \mathbf{Y}$$

از آن‌جایی که ستون‌های $\mathbf{X}$ و $\mathbf{Y}$ برای تضمین اینکه میانگین‌شان صفر باشد استاندارد شده‌اند، ω به‌طور موثری شامل همبستگی هر متغیر با متغیر دیگر می‌شود. پس SIS درحقیقت، متغیر i را انتخاب می‌کند اگر ω_i بین بالاترین مولفه‌های ω باشد.

تعداد متغیرهایی که در مدل انتخاب می‌شود، به اهداف یک محقق بستگی دارد. یک پیشنهاد معقول این است که تا $n-1$ متغیر باقی بماند تا مدل پیشنهادی را داخل قلمرو روش‌های کلاسیک‌تر وارد می‌کند.

برخلاف ظاهر ساده‌اش، نشان داده شده است که SIS ویژگی‌های غربال‌گری مطمئن را داراست. فرض کنید $\mathcal{M}_*$ مدل واقعی مدنظر باشد و فرض کنید $\mathcal{M}_\gamma$ مدل انتخاب شده توسط SIS است. اثبات می‌شود

$$P(\mathcal{M}_* \subset \mathcal{M}_\lambda) \to 1 \quad as \quad n \to \infty \tag{۱۳.۳}$$

برای مشاهده اثبات نتیجه (۱۳.۳) فن و ال‌وی (۲۰۰۸) را ببینید.

[20] Identifiablity
[21] Pruning

SIS مکرر (ISIS)

SIS ممکن است شکست بخورد اگر

- یک متغیر با Y تواما همبسته[۲۲] است (یعنی وقتی در ترکیب با دیگر متغیرها درنظر گرفته شود، روی پاسخ اثر می‌گذارد) اما به‌طور حاشیه‌ای ناهمبسته است (یعنی خودش، خیلی بر پاسخ اثر نخواهد گذاشت). در آن حالت، SIS به متغیر به‌اندازه کافی رتبه نمی‌دهد و به‌سادگی از آن می‌گذرد.

  به‌عنوان یک مثال، یک مدل واقعی شامل $J+1$ متغیر X_1، ...، X_{J+1} را درنظر بگیرید به‌طوری که $\mathrm{Cov}(X_i, X_j) = \rho$. تصور کنید مدل واقعی عبارت است از

$$Y = X_1 + \cdots + X_J - J\rho X_{J+1}.$$

  بدیهی است اگر J بزرگ باشد، X_{j+1} به‌شدت بر Y اثر می‌گذارد. با این حال،

$$\mathrm{Cov}(X_{J+1}, y) = \sum_{i=1}^{J} \mathrm{Cov}(X_{J+1}, X_i) - \mathrm{Cov}(X_{J+1}, J\rho X_{J+1}) = 0$$

  از این رو، SIS این متغیر مهم را در فهرست متغیرهای ذخیره در رتبه آخر قرار خواهد داد!

- یک متغیر ساختگی[۲۳] تنها با توجه به همبستگی‌اش با دیگر متغیرهایی که واقعا با Y همبسته‌اند، با Y همبسته است. این متغیرهای فریبنده، ممکن است بالاتر از دیگر متغیرهای واقعی رتبه بگیرند.

  به‌عنوان یک مثال، مدل واقعی شامل سه متغیر X_0، X_1 و X_2 را درنظر بگیرید به‌طوری که X_0 با دو متغیر دیگر ناهمبسته است. تصور کنید مدل واقعی عبارت است از

$$Y = \rho X_0 + X_1 + X_2$$

  اکنون متغیر چهارم، X_3، را تصور کنید که با X_1 و X_2 همبسته است به‌طوری که $\mathrm{Cov}(X_3, X_1) = \mathrm{Cov}(X_3, X_2) = \rho$ اما با X_0 ناهمبسته است. پس،

$$\mathrm{Cov}(X_0, Y) = \rho$$

$$\mathrm{Cov}(X_3, Y) = \mathrm{Cov}(X_3, X_1 + X_2) = 2\rho$$

  X_3، همبسته ساختگی، با شدت بیشتری از X_0 با Y همبسته است. بنابراین SIS، X_3 را بالاتر از X_0 رتبه خواهد داد.

[22] Jointly correlated

[23] Spurious variable

SIS مکرر (ISIS) روشی است که به‌دنبال غلبه بر این مشکلات طرح شده است.این روش اولین بار در فن و ال‌وی (۲۰۰۸) و یک نسخه بهبودیافته در فن و همکاران (۲۰۰۹) پیشنهاد شده است. در ادامه، دومی بیان می‌شود.

۱. با یک مسئله بعد بسیار بالا شامل p متغیر شروع کنید.

۲. با استفاده از SIS، k_1 تا از این متغیرها را انتخاب کنید. فرض کنید $\mathcal{A}_1 < P$ مجموعه این متغیرها باشد.

۳. از لاسو (یا هر روش مشابه دیگر) برای انتخاب یک زیرمجموعه از این متغیرها استفاده کنید. فرض کنید

- مجموعه نتیجه شده از متغیرها، $\mathcal{M}_1$ باشد.
- بردار ضرایب برازش شده، $\hat{\boldsymbol{\beta}}_{\mathcal{M}_1}$ باشد.
- مولفه‌های زیر_بردار x_i که عضو $\mathcal{M}_1$ هستند را با نماد $x_{i,\mathcal{M}_1}$ نشان دهید.

۴. به هر متغیر j که در مجموعه $\mathcal{M}_1$ انتخاب نشده است، نگاه کنید و کمیت

$$L_j^2 = \min_{\boldsymbol{\beta}_{\mathcal{M}_1}, \beta_j} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, x_{i,\mathcal{M}_1}^\top \boldsymbol{\beta}_{\mathcal{M}_1} + X_{ij}\beta_j)$$

را محاسبه کنید. این کمیت، به‌نظر ترسناک می‌رسد، اما در حقیقت، بسیار ساده است و کارهای زیر را انجام می‌دهد:

- یک مدل شامل همه متغیرها در $\mathcal{M}_1$ و متغیر j درنظر می‌گیرد.
- کمینه $\overline{\text{err}}$ را برای آن مدل پیدا می‌کند.

به‌طور موثری، می‌پرسد «اگر متغیر j به‌همراه آن متغیرها در $\mathcal{M}_1$ کمیت L_j^2 به مدل اضافه شده باشد، چه‌قدر توانسته است $\overline{\text{err}}$ کم شود؟»

۵. متغیرهای j با کمترین مقدار L_j^2 را انتخاب کنید و آن‌ها را در مجموعه $\mathcal{A}_2$ قرار دهید.

۶. اکنون مجموعه $\mathcal{M}_1 \cup \mathcal{A}_2$ را درنظر بگیرید و با استفاده از لاسو (یا هر روش مشابه دیگر) یک زیرمجموعه از این متغیرها انتخاب کنید. مجموعه انتخاب شده از متغیرها را $\mathcal{M}_2$ بنامید.

۷. به گام ۴ بازگردید و به‌جای $\mathcal{M}_1$ از $\mathcal{M}_2$ استفاده کنید.

این الگوریتم می‌تواند آن‌قدر تکرار شود تا $\mathcal{M}_l = \mathcal{M}_{l-1}$ یا به یک مجموعه شامل تعداد از قبل مشخص‌شده‌ای از متغیرهای d دست پیدا کند. فن و همکاران، (۲۰۰۹) $k_1 = \lfloor 2d/3 \rfloor$ را

انتخاب کردند و پس از آن $k_r = d - ||\mathcal{M}_{r-1}||$ تا تضمین کند که ISIS حداقل دو بار تکرار می‌شود.

چگونه ISIS با کاستی‌های SIS مقابله می‌کند؟

- حتی اگر یک متغیر در مرحله اول ISIS رتبه خیلی بالایی نداشته باشد، بسیار محتمل است در مرحله بعدی انتخاب شود، درصورتی‌که با Y تواما همبسته باشد.
- حتی اگر یک متغیر به‌شدت با Y همبسته باشد، آن متغیر انتخاب نخواهد شد مگر اینکه به‌طور قابل توجهی دقت پیشگویی مدل را بهبود دهد. بنابراین در مثالی که مطرح شد، وقتی X_1 و X_2 انتخاب شده‌اند، X_3 نیز بعید است که انتخاب شود زیرا به‌طور قابل توجهی مدل را بهبود نخواهد داد. این موضوع اجازه می‌دهد تا X_0 انتخاب شود.

انواع مختلف ISIS

تلاش‌های متعددی برای بهبود عملکرد ISIS انجام شده‌اند. به‌طور خیلی خلاصه دو مورد از آن‌ها، در ادامه بیان می‌شوند:

- فن و ال‌وی (۲۰۰۸) پیشنهاد کردند که تبدیل متغیرها به‌عنوان راهی برای رویارویی با همبستگی استفاده شود. برای مثال، وزن‌های w_1، w_2 و w_3 مربوط به ۲، ۹ و ۱۸ ساله‌ها باشند. بدیهی است بین آن‌ها همبستگی مثبت وجود دارد. با در نظر گرفتن متغیرهای w_1، $w_2 - w_1$ و $w_3 - w_2$ می‌توان همبستگی را به‌طور قابل ملاحظه‌ای تضعیف کرد.
- فن و همکاران (۲۰۰۹) روش زیر را پیشنهاد کردند:
  - n داده را به‌طور تصادفی به دو قسمت تقسیم کنید.
  - به‌طور جداگانه SIS یا ISIS را برای داده‌ها در هر قسمت اجرا کنید که دو مجموعه از متغیرهای $\mathcal{A}_1$ و $\mathcal{A}_2$ را نتیجه خواهد داد.
  - هر دو این مجموعه‌ها در ویژگی غربال‌گری مطمئن (رابطه (۱۳.۳)) صدق می‌کنند و از این رو، شامل متغیرهای زیادی خواهند بود که به‌درستی در مدل هستند. با این حال، آن‌ها همچنین شامل متغیرهای زیادی که در مدل موردنظر نیستند، نیز خواهند بود (به‌عبارت دیگر، نرخ کشف اشتباه [۲۴] (FDR) برای این مجموعه‌ها بالا خواهد بود).
  - مجموعه

$$\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$$

[24] False Discovery Rate (FDR)

را بسازید. این مجموعه نیز در ویژگی غربال‌گری مطمئن صدق می‌کند اما هچنان خیلی کمتر شامل مثبت‌های اشتباه [۲۵] خواهد بود.

درباره روش SIS، خواننده علاقه‌مند به کاظمی (۱۳۹۷) ارجاع داده می‌شود.

رویارویی با ناتوانی‌ها در شرایط نظم ـ لاسو تصادفی شده

ژو و یو (۲۰۰۶) نشان دادند که لاسو، فقط در صورتی به‌طور سازگار متغیرها را انتخاب می‌کند که مدل در شرط نمایش‌ناپذیری صدق کند. ژو (۲۰۰۶) یک راه حل برای این مسئله در لاسو سازوار، که یک فرآیند دو مرحله‌ای پاست، پیشنهاد کرد. در اینجا، بنا به مینشاوسن و بولمن (۲۰۱۰) یک الگوریتم جایگزین به نام لاسو تصادفی‌شده، در نظر گرفته می‌شود. این روش برخلاف سادگی‌اش، حتی اگر شرط نمایش‌ناپذیری نقض شده باشد، برای انتخاب متغیر، سازگار است.

ایده اصلی لاسو تصادفی‌شده، درحقیقت، تغییر پارامتر تنظیم‌کننده λ برای هر مولفه β است.

تعریف ۱.۲.۳ (لاسو تصادفی شده). یک مقدار $\alpha \in (0, 1]$ (ضعف الگوریتم) انتخاب کنید و فرض کنید W_k متغیرهای مستقل و هم‌توزیع در $[\alpha, 1]$، برای $k = 1, \ldots, p$ باشند. برآوردگر لاسو تصادفی شده عبارت است از

$$\hat{\boldsymbol{\beta}}^{\mathrm{LASSO},\lambda,\mathrm{W}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k}$$

البته، امید به اینکه این انحراف تصادفی به‌طور یکنواخت به یک بهبود منجر خواهد شد، منطقی نیست ـ و در واقع، اگر یک بار به‌کار گرفته شود، لاسو تصادفی‌شده خیلی مفید نیست. با این حال، به‌کارگیری چندین باره لاسو تصادفی‌شده و جستجو برای متغیرهایی که اغلب انتخاب شده‌اند، به‌نظر می‌رسد یک فرآیند بسیار پرقدرت است.

[25] False positive

فصل ۴

روش‌های انتخاب متغیر گروهی

همان‌طور که در فصل‌های پیشین اشاره شد، روش‌های انتخاب متغیر کلاسیک بر اساس معیارهایی مانند C_pـ مالو (مالو، ۱۹۷۳)، AIC (آکائیک، ۱۹۷۴) و BIC (شوارتز، ۱۹۷۸) از روش‌های اخیر بر اساس منظم‌سازی، می‌توان به بریج (فرانک و فریدمن، ۱۹۹۶)، لاسو (تیبشیرانی، ۱۹۹۶)، SCAD (فن و لی، ۲۰۰۱؛ فن و پنگ، ۲۰۰۴) و الاستیک‌نت (ژو و هیستی، ۲۰۰۵) اشاره کرد. این روش‌ها به‌منظور انتخاب متغیرها به‌صورت جداگانه (خارج از گروه) طراحی شده‌اند.

در برخی از مطالعات مانند ANOVA لزوم انتخاب گروهی از متغیرها به جای انتخاب متغیرها به‌صورت تکی وجود دارد. در این مسائل کاربردی، یک بردار پارامتر بعد بالا $\boldsymbol{\beta}$ در یک مدل رگرسیون، از گروه‌های $\mathcal{G}_1$، $\mathcal{G}_2$، ... و $\mathcal{G}_q$ که یک افزار از مجموعه اندیس‌های $\{1,2,\ldots,p\}$ هستند، ساخته شده است. به‌عبارت دیگر،

$$\cup_{j=1}^{q} \mathcal{G}_j = \{1,2,\ldots,p\} \qquad \text{و} \qquad \mathcal{G}_j \cap \mathcal{G}_k = \phi \; j \neq k.$$

بنابراین، بردار پارامتر $\boldsymbol{\beta}$، به‌صورت

$$\boldsymbol{\beta}_{\mathcal{G}} = (\boldsymbol{\beta}_{\mathcal{G}_1}, \ldots, \boldsymbol{\beta}_{\mathcal{G}_q})^\top \qquad \boldsymbol{\beta}_{\mathcal{G}_r} = \{\boldsymbol{\beta}_r; r \in \mathcal{G}_r\} \tag{۱.۴}$$

است. یک خانواده مهم از مثالی برای چنین ساختار گروهی، متغیرهای موهومی هستند. به‌عنوان مثال، متغیر پاسخ حقیقی‌ـ مقدار Y و p متغیر پیش‌گوی X_1، X_2، ... و X_p در نظر بگیرید که هر X_j به مجموعه $\mathcal{X}$ تعلق دارد ($X_j \in \mathcal{X}$) و X_jها ۴ـ سطحی هستند که با یکی

از اعضای $\{۱,۲,۳,۴\}$ کدگذاری شده‌اند. بنابراین، برای مدل‌بندی اثر اصلی، به ۳ پارامتر نیاز است. اگر اثر متقابل مرتبه اول در مدل وجود داشته باشد، تعداد پارامترها، ۹ تا است و به‌همین ترتیب. با درنظر گرفتن نمادگذاری بیان شده برای بردار $\boldsymbol{\beta}$، ساختار گروهی به‌صورت زیر است: اثر اصلی متغیر $X_۱$ متناظر با $\boldsymbol{\beta}_{\mathcal{G}_۱}$ با $|\boldsymbol{\beta}_{\mathcal{G}_۱}| = ۳$ است و به همین ترتیب، اثر اصلی سایر متغیرهای X_j متناظر با $\boldsymbol{\beta}_{\mathcal{G}_j}$ با $|\boldsymbol{\beta}_{\mathcal{G}_j}| = ۳$، به‌ازای $j = ۱, \ldots, p$ است. علاوه بر این، یک اثرمتقابل از $X_۱$ و $X_۲$ متناظر با $\boldsymbol{\beta}_{\mathcal{G}_{p+۱}}$ با $|\boldsymbol{\beta}_{\mathcal{G}_{p+۱}}| = ۹$ است.

گروه‌بندی می‌تواند به مدل بر اساس دانش و اطلاعات قبلی که از نظر علمی، معنی‌دار و منطقی هستند، اضافه شود. به‌عنوان مثال، در تحلیل ژن‌ها، ژن‌های متعلق به مسیر بیولوژیکی مشابه، می‌توانند به‌عنوان یک گروه در نظر شوند. در مطالعات ژنتیک، نشان‌گرهای ژنتیکی[۱] از ژن‌های مشابه، می‌توانند به‌عنوان یک گروه در نظر گرفت. در نظر گرفتن ساختارهای گروهی در تحلیل چنین داده‌هایی، مطلوب است.

بنا به موقعیت، متغیرهای درون هر گروه ممکن است از دیدگاه علمی، معنی‌دار باشند یا نباشند. اگر آن‌ها، به‌تنهایی، معنی‌دار نباشند، معمولا علاقه‌مندی به انتخاب متغیر تکی وجود ندارد و هدف محقق، کاملا، معطوف به انتخاب گروهی می‌شود. اما اگر متغیرها به‌تنهایی معنی‌دار باشند، آن‌گاه باید در کنار انتخاب گروهی متغیرها، متغیرهای بااهمیت به‌طور تک به تک نیز انتخاب که از آن به انتخاب دوـسطحی تعبیر می‌شود. برای مثال، اگر یک متغیر پیوسته بر اساس یک مجموعه از پایه‌ها یا یک مجموعه گسسته نمایش داده شود، متغیرها در خارج از گروه، از یک ساختار مصنوعی پیروی می‌کنند و در این حالت، انتخاب اعضای تکی گروه‌ها، معمولا، مورد علاقه محقق نیست. اما، در مثال‌های مربوط به ژن‌ها و نشان‌گرهای ژنتیکی، انتخاب تکی متغیرها درست به اندازه انتخاب گروهی متغیرها، اهمیت دارد. برای مثال، فرض کنید $\mathcal{G}_۱$، ... و $\mathcal{G}_q$ نقش‌های متفاوت بیولوژیکی را برای ژن‌های گروه‌بندی شده نشان می‌دهند و پاسخ y، یک بیماری خاص باشد. یافتن ارتباط بین ژن‌ها و ژن‌های گروه‌بندی شده با بیماری، هدف موردنظر است. در سایر مثال‌ها، مانند یک گروه از تابع‌های نشان‌گر[۲] برای یک متغیر گروهی، انتخاب تکی متغیرها، به مسئله مورد بررسی بستگی دارد. در عمل، ساختارهای گروهی معمولا در علوم زیستی، مهندسی و روش‌های آماری در مطالعات علوم اقتصادی مشاهده می‌شوند.

لاسو (تیبشیرانی، ۱۹۹۶) یک رهیافت کارآمد برای مسئله «p بزرگ، n کوچک» است. می‌توان لاسو را در حالت «گروهی» استفاده کرد. به‌عبارت دیگر، بر اساس نماد گذاری بیان شده، لاسو در این حالت، تابع

$$\|\boldsymbol{y} - \mathbf{X}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}}\|_۲^۲ + \lambda \sum_{i=۱}^{q} |\boldsymbol{\beta}_{\mathcal{G}_i}|$$

را کمینه می‌کند که q تعداد گروه‌ها، $\boldsymbol{\beta}_{\mathcal{G}} = (\boldsymbol{\beta}_{\mathcal{G}_۱}, \ldots, \boldsymbol{\beta}_{\mathcal{G}_q})^\top$ و $X_{\mathcal{G}}$ ماتریسی است که بر اساس

[1] Genetic markers

[2] Indicator function

گروه‌ها، دوباره نویسی شده است. بدیهی است به‌ازای مشاهده iام، $\mathbf{X}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}} = \sum_{k=1}^{q}\sum_{j=1}^{p_i}\boldsymbol{\beta}_{kj}\mathbf{X}_{i(kj)}$ و اگر فرض کنید $\boldsymbol{\beta}_{\mathcal{G}_i} = \{\boldsymbol{\beta}_r, r \in \mathcal{G}_i\} = (\beta_1, \ldots, \beta_{p_i})^\top$، آن‌گاه $|\boldsymbol{\beta}_{\mathcal{G}_i}| = \sum_{j=1}^{p_i}|\beta_j|$. هم‌چنین $\lambda \geq 0$ یک پارامتر منظم‌سازی است.

لاسو برای متغیرهایی که به تنهایی رفتار می‌کنند، می‌تواند به‌خوبی عمل کند. وقتی متغیرها، گروه‌بندی شده باشند، لاسو نمی‌تواند خوب کار کند. به‌عنوان مثال، فرض کنید k امین گروه بی‌اهمیت باشد، لاسو، فقط ضریب متناظر با گروه k ام را صفر می‌کند و از آن‌جایی که k امین گروه، بی‌اهمیت است، تمام ضرایب این گروه باید صفر باشند. لاسو، در واقع با گروه‌ها مشابه متغیرهای تکی برخورد می‌کند و انتخاب را بر اساس قدرت متغیرهای تکی انجام می‌دهد تا بر اساس قدرت گروه‌ها. بنابراین، برای حالت‌های گروهی، بهتر است اعضای گروه، با یکدیگر منقبض و انتخاب شوند.

در این راستا، مطالعات بسیاری به‌منظور انتخاب گروه‌های مهم از متغیرها انجام گرفته‌اند. یوآن و لین (۲۰۰۶) روش‌های لاسوی گروهی و LARS گروهی را معرفی کردند. لاسو گروهی، یک تعمیم ساده از لاسو است که در آن یک نرم L_2 از ضرایب با گروهی از متغیرها ارتباط دارند که مولفه تابع جریمه است. این روش، قادر نیست تا به‌طور همزمان انتخاب متغیر گروهی و مجزا را انجام دهد. حتی این موضوع، در بسیاری از مسائل، مطلوب نیست. از این رو، هوانگ و همکاران (۲۰۰۹)، مسئله انتخاب دوسطحی را در نظر گرفتند و روش بریج گروهی را پیشنهاد کردند.

۱.۴ لاسو گروهی

در برآورد مدل‌ها با ساختار گروهی برای بردار پارامترها، اغلب هدف مطلوب، ایجاد تنکی در سطح گروه‌ها است. در این حالت، همه مولفه‌های $\hat{\beta}_{\mathcal{G}_j}$ باید صفر یا همه مولفه‌ها باید غیرصفر باشند. این هدف با استفاده از جریمه لاسو گروهی (یوان و لین، ۲۰۰۶) به‌صورت

$$\lambda\sum_{j=1}^{q} m_j \left\|\boldsymbol{\beta}_{\mathcal{G}_j}\right\|_2 \qquad (۲.۴)$$

دست‌یافتنی است که در آن $\|\cdot\|_2$ نرم استاندارد اقلیدسی (L_2) است. ضریب m_j، تعادل را در حالت‌هایی ایجاد می‌کند که گروه‌ها اندازه‌های خیلی متفاوتی دارند و معمولا به‌صورت

$$m_j = \sqrt{T_j}$$

در نظر گرفته می‌شود که در آن T_j اندازه $|\boldsymbol{\beta}_{\mathcal{G}_j}|$ را نشان می‌دهد. این تابع جریمه، حد واسط میان جریمه L_1 در رگرسیون لاسو و جریمه L_2 در رگرسیون ریج است. از آنجا که نرم اقلیدسی یک بردار $\boldsymbol{\beta}_{\mathcal{G}}$ تنها زمانی صفر است که همه مؤلفه‌های آن صفر باشند، این رویه، تُنُکی را در هر دو سطح گروهی و تک تک متغیرها تقویت می‌کند. به این معنی که برای بعضی مقادیر

λ، پارامتر تنظیم‌کننده مشخص λ می‌تواند به‌عنوان مثال، ضریب بردار $\boldsymbol{\beta}_{\mathcal{G}_j}$ را برابر صفر قرار دهد. بنابراین گروه jام از مدل برازش شده حذف می‌شود. اگر اندازه همه گروه‌ها برابر با ۱ باشد، این روش به لاسو تبدیل می‌شود.

لاسوی گروهی، تابع

$$||\boldsymbol{y} - \mathbf{X}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}}||_2^2 + \lambda \sum_{j=1}^{q} m_j \left\|\boldsymbol{\beta}_{\mathcal{G}_j}\right\|_2$$

را کمینه می‌کند. با توجه به اینکه، این تابع، محدب است، هر کمینه موضعی، یک کمینه مطلق نیز است که یوآن و لین (۲۰۰۶) راه‌حلی برای یافتن پاسخ آن ارائه دادند.

البته، بدون از دست دادن کلیت مسئله، بهتر است متغیرهای پیش‌گو و پاسخ مرکزی شوند. در این صورت، هر گروه از متغیرها، مثل یک واحد، رفتار می‌کنند و در زمان برآورد ضرایب رگرسیونی و انتخاب متغیرهای با اهمیت، مزیت ساختار گروهی در این مدل‌ها مشخص می‌شود.

رهیافت لاسو گروهی، ابتدا توسط باکین (۱۹۹۹) به‌عنوان تعمیمی از لاسو برای انتخاب گروه‌های متغیرها پیشنهاد شد که برای آن یک الگوریتم محاسباتی نیز ارائه کرد. یوان و لین (۲۰۰۶) کار باکین را توسعه دادند. اخیرا امکان هم‌پوشانی گروه‌های پیشگوها (ژائو و همکاران، ۲۰۰۹) و ارتباط با روش‌هایی برای برازش مدل‌های افزودنی تُنُک وجود دارد (لین و ژانگ، ۲۰۰۶؛ راویکومار و همکاران، ۲۰۰۸). ژانگ و هوانگ (۲۰۰۸) ویژگی‌های مجانبی این برآوردگر را بررسی کرده‌اند.

یک سوال ساده درباره لاسو گروهی این است که تحت چه شرایطی، بهتر از برآوردگر لاسو عمل می‌کند. این سوال توسط هوانگ و ژانگ (۲۰۱۰) پاسخ داده شده است که مفهوم تنکی شدید گروهی را معرفی کرد. آن‌ها نشان دادند که لاسو گروهی تحت تنکی گروهی شدید و شرایط خاص دیگری، مانند شرط مقدار ویژه تنکی گروهی، بر لاسو برتری دارد.

بریج گروهی

بریهینی و هوانگ (۲۰۰۹) مسئله‌های رگرسیونی را در نظر گرفتند که متغیرهای کمکی می‌توانند گروه‌بندی شده باشند. آن‌ها علاقه‌مند بودند که انتخاب گروه‌های مهم را به‌خوبی تشخیص اعضای مهم این گروه‌ها، انجام دهند و آن را به‌عنوان انتخاب دوسطحی تعریف کردند. آن‌ها ذکر کردند تابع جریمه لاسو گروهی (یوآن و لین، ۲۰۰۶) از نرم‌های L_2 گروه‌ها ساخته شده است که این نرم‌ها باعث می‌شوند انقباضی مانند رگرسیون ریج به درون گروه‌ها اعمال می‌کند درحالی که در بین گروه‌ها، دارای اثر تُنُکی است. لاسو گروهی، انتخاب متغیر را در سطح گروهی انجام می‌دهد اما نمی‌تواند داده‌ها را به‌طور تک به تک، انتخاب کند.

بریج گروهی (هوانگ و همکاران، ۲۰۰۹) جریمه بریج نرم L_1 را درون گروه‌ها به‌کار می‌گیرد

و با ایجاد تنکی در هر دو سطح تکی و گروهی انتخاب متغیر را اجرا می‌کند. بریج گروهی، تابع

$$||\boldsymbol{y} - \mathbf{X}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}}||_{۲}^{۲} + \lambda \sum_{j=۱}^{q} c_j \left\|\boldsymbol{\beta}_{\mathcal{G}_j}\right\|_{۱}^{\gamma} \qquad (۳.۴)$$

را کمینه می‌کند که $\lambda > ۰$، پارامتر جریمه و m_j ها ثابت‌هایی هستند که بعدهای متفاوت بردارهای $\boldsymbol{\beta}_{\mathcal{G}_j}$ را تعدیل می‌کند. یک انتخاب ساده می‌تواند $c_j \propto |\mathcal{G}_j|^{۱-\gamma}$ باشد. در این جا، گروه‌های $\mathcal{G}_j$ می‌توانند هم‌پوشانی داشته باشند و اجتماع آن‌ها اجازه دارد که یک زیرمجموعه مناسب از همه مجموعه‌ها باشد و لزوما نباید با $\cup_{i=۱}^{q}\mathcal{G}_i$ برابر باشد. بنابراین، متغیرهایی که در $\cup_{i=۱}^{q}\mathcal{G}_i$ هستند، اما در آن زیر مجموعه مناسب نیست، جریمه نمی‌شوند. از آن‌جایی‌که ساختار گروهی، فقط در تابع جریمه نمود دارد، هم‌پوشانی فقط در قسمت جریمه این تابع امکان‌پذیر است. برای مثال، اگر یک متغیر به دو گروه تعلق داشته باشد، آن‌گاه ضریب متناظر با این متغیر، در جریمه دو بار ظاهر شده است.

هنگامی که به‌ازای $۱ \leq q \leq ۱$، $|\mathcal{G}_i| = ۱$؛ آن‌گاه بریج گروهی به بریج معمولی تبدیل می‌شود. وقتی $\gamma = ۱$، جریمه بریج همان جریمه لاسو خواهد یود که می‌تواند انتخاب متغیر را در سطح تکی انجام دهد. به‌ازای $۰ < \gamma < ۱$، معیار بریج گروهی (۳.۴) می‌تواند به‌طور همزمان، متغیرها را در دو سطح تکی و گروهی، انتخاب کند.

۲.۴ ساختار روش‌های جریمه‌شده گروهی

اثر جریمه بر ضرایب، وابسته به مشتق جریمه هستند. جریمه‌ها صورتی مانند $\lambda\beta^{\gamma}$ دارند. در رگرسیون ریج $\gamma = ۲$، بنابراین میزان جریمه آن با افزایش β افزایش می‌یابد. جریمه ریج، مقدار خیلی کمی (تقریبا نزدیک صفر) ضرایب نزدیک به صفر را جریمه می‌کند اما به ضرایبی که مقدار بزرگی دارند، جریمه‌های بزرگی اعمال می‌کند. در لاسو $\gamma = ۱$؛ میزان جریمه ثابت است. اگر $\gamma = \frac{۱}{۲}$ میزان جریمه به میزان زیادی نزدیک به صفر است اما با بزرگتر شدن β، کاهش می‌یابد.

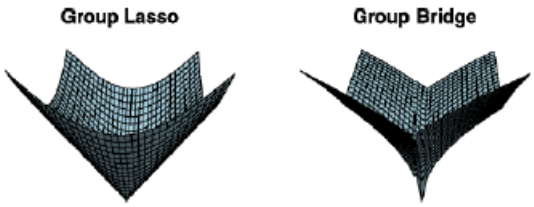
در حالتی که داده‌ها در شرایط نرمال‌سازی صدق کنند، لاسو گروهی، تابع هدف (۴.۴) را کمینه می‌کند

$$\frac{۱}{۲n}||\mathbf{y} - \mathbf{X}\beta||^{۲} + \lambda \sum_{j=۱}^{q} \sqrt{p_j}||\beta_j||, \qquad (۴.۴)$$

و $\boldsymbol{\beta}$ کمینه آن است. برآورد بریج گروهی، با کمینه کردن تابع

$$\frac{۱}{۲n}||\mathbf{y} - \mathbf{X}\beta||^{۲} + \lambda \sum_{j=۱}^{q} p_j^{\gamma}||\beta_j||_{۱}^{\gamma}, \qquad (۵.۴)$$

نتیجه می‌شود. در ادامه، فرض می‌شود بریج گروهی دارای $\gamma = ۱/۲$ است.



شکل ۱.۴: جریمه‌های اعمال شده بر یک گروه دو متغیره با روش‌های لاسو گروهی و بریج گروهی

شکل ۱.۴ جریمه‌های اعمال شده بر یک گروه دو متغیره با روش‌های لاسو گروهی و بریج گروهی را نشان می‌دهد. توجه کنید که جایی‌که جریمه به یک نقطه یا لبه می‌رسد، احتمال این وجود دارد که پاسخ مقداری تُنُک باشد؛ همه جریمه‌هایی که به نقطه صفر می‌رسند، تُنُکی در سطح گروهی را ایجاد می‌کنند، اما فقط بریج گروهی و انتخاب دو سطحی را اجازه می‌دهد.

الگوریتم کاهش مختصات موضعی برای لاسو گروهی و بریج گروهی

الگوریتم‌های کاهش مختصات[۳]، یک تابع هدف را نسبت به یک پارامتر تکی در هر بار از یک چرخه تکراری درمیان همه پارامترها بهینه می‌کند تا به همگرایی برسد. اخیرا، قدرت الگوریتم‌های کاهش مختصات برای بهینه‌سازی مسائل رگرسیون جریمه‌شده مورد توجه بسیاری از محققان قرار گرفته است.

روند الگوریتم کاهش مختصات موضعی[۴] (LCD) گروهی به‌صورت زیر است:

۱. یک برآورد اولیه، $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(0)}$ انتخاب کنید.

۲. متغیرهای پیش‌گو را به صورت زیر به‌روزرسانی کنید.

(a) $\tilde{b}_0$ را به‌روزرسانی کنید.

(b) برای $j \in 1, \ldots, J$ ، $\tilde{\beta}_j$ را به‌روزرسانی کنید.

۳. مرحله ۲ را تا رسیدن به همگرایی تکرار کنید.

در ابتدا بریهینی و هوانگ (۲۰۰۹)، عرض از مبدأ در مرحله ۲a را به‌روزرسانی کردند. باقی‌مانده‌های جزئی[۵] برای به‌روزرسانی $\tilde{\beta}_0$، $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{X}_{-0}\tilde{\boldsymbol{\beta}}_{-0}$ است که در آن $\mathbf{X}_{-0}$، ماتریس $\mathbf{X}$ بدون ستون عرض از مبدا است (دقت کنید، وقتی‌داده‌ها در شرایط نرمال‌سازی صدق کنند،

[3] Coordinate descent algorithm
[4] Local Coordinate Descent (LCD)
[5] Partial residuals

عرض از مبدا در مدل وجود ندارد). از این رو، مقدار به‌روز شده $\tilde{\boldsymbol{\beta}}_{\circ}$ جواب رگرسیون خطی ساده است، بنابراین،

$$\tilde{\boldsymbol{\beta}}_{\circ} \longleftarrow \frac{\mathbf{X}_{\circ}^{\top}\tilde{\boldsymbol{y}}}{\mathbf{X}_{\circ}^{\top}\mathbf{X}_{\circ}} = \frac{1}{n}\mathbf{X}_{\circ}^{\top}\tilde{\boldsymbol{y}}.$$

آن‌ها همچنین یک رهیافت هم‌ارز با باقی‌مانده‌ها برای به‌روزرسانی $\tilde{\boldsymbol{\beta}}_{\circ}$ استفاده کردند که کارایی بیشتری دارد. آن‌ها $\tilde{\boldsymbol{\beta}}_{\circ}$ را با بهره گرفتن از باقی‌مانده‌های متداول $\tilde{\boldsymbol{r}} = \boldsymbol{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ به‌روزرسانی کردند. از آن‌جایی که $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{r}} + \mathbf{X}_{\circ}\tilde{\boldsymbol{\beta}}_{\circ}$، به‌روزرسانی به‌صورت

$$\tilde{\boldsymbol{\beta}}_{\circ} \longleftarrow \frac{1}{n}\mathbf{X}_{\circ}^{\top}\tilde{\boldsymbol{r}} + \tilde{\boldsymbol{\beta}}_{\circ} \qquad (۶.۴)$$

می‌شود. به‌روزرسانی $\tilde{\boldsymbol{\beta}}_{\circ}$ در این روش به تنها $2n$ کار نیاز دارد: n کار برای محاسبه $\mathbf{X}_{\circ}^{\top}\tilde{\boldsymbol{r}}$ و n کار برای به‌روزرسانی $\tilde{\boldsymbol{r}}$. در مقابل، به‌دست آوردن $\tilde{\boldsymbol{y}}$ به $n \times (p-1)$ کار نیاز دارد. در ضمن، در روش بهینه‌سازی تکراری وزنی[۶]، مرحله به‌روزرسانی عبارت است از

$$\frac{\tilde{\boldsymbol{\beta}}_{\circ}\mathbf{X}_{\circ}^{\top}\boldsymbol{W}\tilde{\boldsymbol{r}}}{\mathbf{X}_{\circ}^{\top}\boldsymbol{W}\mathbf{X}_{\circ}} + \tilde{\boldsymbol{\beta}}_{\circ} \qquad (۷.۴)$$

که نیاز به $3n$ کار دارد. به‌روزرسانی $\tilde{\beta}_j$ در گام $2b$ به نوع جریمه‌ها بستگی دارد. در ادامه، مراحل به‌روزرسانی برای جریمه‌های مختلف گروهی، لاسو گروهی و بریج گروهی، مورد بحث قرار می‌گیرد.

بریج گروهی

الگوریتم کاهش مختصات موضعی برای بریج گروهی به‌صورت زیر است:

وقتی جریمه $\lambda|\boldsymbol{\beta}|$ به یک پارامتر اعمال شود، $\lambda|\boldsymbol{\beta}|$، پاسخ لاسو (تیبشیرانی، ۱۹۹۶) عبارت است از

$$\boldsymbol{\beta} = \frac{S(\frac{1}{n\mathbf{X}^{\top}}\boldsymbol{y}, \lambda)}{\frac{1}{n\mathbf{X}^{\top}\mathbf{X}}} = S(\frac{1}{nX^{\top}}\boldsymbol{y}, \lambda),$$

که در آن $S(z,c)$ عملگر آستانه نَرم[۷] (دونهو و جانستون، ۱۹۹۴) برای c مثبت به‌صورت

$$S(z,c) = \begin{cases} z-c, & z > c, \\ 0, & |z| \le c, \\ z+c, & z < -c \end{cases}.$$

تعریف می‌شود. با در نظر گرفتن تقریب سری تیلور مرتبه اول پیرامون $\tilde{\boldsymbol{\beta}}_j$، جریمه به‌عنوان تابعی از β_{jk} تقریبا با $\tilde{\lambda}_{jk}|\beta_{jk}|$ متناسب است، که در آن

$$\tilde{\lambda}_{jk} = \lambda_{\gamma} K_j^{\gamma} ||\tilde{\beta}_j||^{\gamma-1}. \qquad (۸.۴)$$

[6] Iteratively reweighted optimization

[7] Soft threshold

از این رو در ناحیه موضعی که در آن جریمه توسط یک تابع خطی به‌خوبی تقریب زده شده است، گام $2b$ از گام‌های به‌روزرسانی ساده مبتنی بر برش آستانه نَرم برای $\tilde{\lambda}_{jk}$ $k \in 1,\ldots,p_j$ تشکیل می‌شود. به‌عبارت دیگر،

$$\tilde{\beta}_{jk} \longleftarrow S(\frac{1}{n\mathbf{x}_{jk}^{\top}\tilde{\boldsymbol{r}}} + \tilde{\beta}_{jk}, \tilde{\lambda}_{jk}) \tag{۹.۴}$$

یا بر اساس وزن‌ها،

$$\tilde{\beta}_{jk} \longleftarrow \frac{S(\frac{1}{n\mathbf{x}_{jk}^{\top}\boldsymbol{W}\tilde{\boldsymbol{r}}} + \frac{1}{n\mathbf{x}_{jk}^{\top}\boldsymbol{W}\mathbf{x}_{jk}\tilde{\boldsymbol{\beta}}_{jk},\tilde{\lambda}_{jk})}}{\frac{1}{n\mathbf{x}_{jk}^{\top}\boldsymbol{W}\mathbf{x}_{jk}}}. \tag{۱۰.۴}$$

مشکل در بریج گروهی این است که جریمه بریج همه‌جا مشتق‌پذیر نیست و $\tilde{\lambda}_{jk}$ در $\tilde{\beta}_j = 0$ برای $y < 1$ تعریف نشده است. صفر هم، خودش مشکل اساسی با جریمه درست می‌کند. برای هر مقدار مثبت λ، صفر یک کمینه موضعی برای جریمه بریج گروهی است و از این‌رو، بهینه‌سازی را پیچیده می‌کند.

برای فائق آمدن بر این مشکل، آن‌ها انتخاب کردند که با مقداری اولیه بسیار دور از صفر شروع کنند. اگر $\tilde{\beta}_j$ در هر نقطه‌ای در حین تکرار به صفر برسد، بعد از آن، $\tilde{\beta}_j$ را در صفر نگاه می‌دارند. اشکالی که به وجود می‌آید این است که گروه‌هایی که در واقع غیرصفر هستند،و قتی جواب همگرا می‌شود، از مدل حذف می‌شوند. رهیافت‌های دیگری برای مقابله با این مشکل وجود دارند، مانند اضافه کردن یک ثابت کوچک به $\tilde{\beta}_j$ در (۸.۴)، اگرچه، این از بهره‌جستن الگوریتم از تُنُکی ممانعت می‌کند و به‌شدت کارایی محاسباتی را برای مسائل تُنُک بزرگ کاهش می‌دهد.

لاسو گروهی

به‌روزرسانی در لاسو گروهی با توجه به اینکه ویژگی‌هایش که متغیرهای گروهی یا همه باهم صفر می‌شوند یا اصلا صفر نمی‌شوند، بسیار پیچیده‌تر است. بریهینی و هوانگ (۲۰۰۹) انتخاب کردند که $\tilde{\beta}_j$ گام $2b$ را در دو گام به‌روزرسانی کنند: (۱) چک شود که آیا $\tilde{\boldsymbol{\beta}}_j = 0$، (۲) اگر $\tilde{\boldsymbol{\beta}}_j \neq 0$ آن‌گاه برای $k \in 1,\ldots,p_j$، $\tilde{\boldsymbol{\beta}}_{jk}$ را به‌روزرسانی کنند. گام (۱) تحت این شرط اجرا می‌شود که $\tilde{\boldsymbol{\beta}}_j \neq 0$ اگر و تنها اگر

$$\frac{1}{n}||\mathbf{X}_j^{\top}\tilde{\boldsymbol{r}} + \mathbf{X}_j^{\top}\mathbf{X}_j\tilde{\boldsymbol{\beta}}_j|| > \sqrt{p_j}\lambda \tag{۱۱.۴}$$

شرط‌های بالا شرط‌های کاروش‌ـ کوهن‌ـ تاکر[8] (KKT) برای این مسئله هستند و اولین بار توسط یوان و لین (۲۰۰۶) ذکر شدند. اگر β_j نتواند بدون افزایش جریمه بیش‌تر از حرکتی که برازش را بهبود دهد در هرجهتی به‌دور از صفر حرکت کند، پس صفر کمینه موضعی است. از آن‌جا که جریمه لاسو گروهی محدب است، صفر هم کمینه‌کننده یکتای سراسری است.

[8] Karush-Kuhn-Tacker

اگر این شرط برقرار نباشد، پس آن‌ها $\tilde{\beta}_j = ۰$ را قرار می‌دهند و می‌گذرند. در غیر این صورت، آن‌ها یک تقریب موضعی برای جریمه می‌گیرند و اعضای گروه j را به‌روزرسانی می‌کنند. با این حال، به‌جای تقریب زدن جریمه به‌عنوان تابعی از $|\tilde{\beta}_{jk}|$، می‌توانند با درنظر گرفتن جریمه به‌عنوان تابعی از $\tilde{\beta}_{jk}^۲$ تقریب بهتری به‌دست آورند. بنابراین جریمه اعمال شده بر β_{jk} می‌تواند با $\tilde{\lambda}_{jk}\tilde{\beta}_{jk}^۲/۲$ تقریب زده شود، که در آن

$$\tilde{\lambda}_{jk} = \frac{\lambda\sqrt{p_j}}{||\tilde{\boldsymbol{\beta}}_j||} \tag{۱۲.۴}$$

این رهیافت یک گام به‌روزرسانی انقباضی را به‌جای گام آستانه‌ـنَرم به‌دست می‌دهد

$$\tilde{\beta}_{jk} \longleftarrow \frac{\frac{۱}{n\mathbf{x}_{jk}^\top \tilde{\boldsymbol{r}}} + \tilde{\beta}_{jk}}{۱ + \tilde{\lambda}_{jk}} \tag{۱۳.۴}$$

یا برای بهینه‌سازی وزنی

$$\tilde{\beta}_{jk} \longleftarrow \frac{\frac{۱}{n\mathbf{X}_{jk}^\top \boldsymbol{W}\tilde{\boldsymbol{r}}} + \frac{۱}{n\mathbf{x}_{jk}^\top \boldsymbol{W}\mathbf{x}_{jk}\tilde{\boldsymbol{\beta}}_{jk}}}{\frac{۱}{n\mathbf{x}_{jk}^\top \boldsymbol{W}\mathbf{x}_{jk}} + \tilde{\lambda}_{jk}} \tag{۱۴.۴}$$

توجه کنید که همانند (۸.۴)، (۱۲.۴) نیز در صفر تعریف نشده است. با این حال، این تنها یک الگوریتم جزئی نامناسب در لاسو گروهی است. جریمه مشتق‌پذیر است اما با مشتقات جزئی که دارای صورت متفاوتی در صفر هستند. می‌توان از این مشکل با افزودن یک مقدار مثبت کمی δ به مخرج در (۱۰.۴) ممانعت کرد.

۳.۴ لاسو گروهی برای رگرسیون لجستیک

فرض کنید مجموعه داده‌های $\{(\mathbf{x}_i, y_i); i = ۱, \ldots, n\}$ در دسترس هستند که بردار p بعدی $\mathbf{x}_i$ از q پیش‌گو هستند. $y_i \in \{۰, ۱\}$ یک پاسخ دو تایی است. پیش‌گوها هم می‌توانند پیوسته و هم می‌توانند رسته‌ای باشند. فرض کنید df_p درجه آزادی pامین پیش‌گو باشد. برای مثال، اثر اصلی یک عامل با ۴ سطح، دارای $df = ۳$ است و یک پیش‌گوی پیوسته فقط $df = ۱$ دارد.

رگرسیون لجستیک خطی، احتمال $p_i = P(Y = ۱|\mathbf{x}_i)$ را به‌صورت

$$\eta(\mathbf{x}_i) = \log(\frac{p_i}{۱ - p_i}) = \beta_۰ + \mathbf{x}_i^\top \boldsymbol{\beta} \tag{۱۵.۴}$$

مدل می‌کند که $\beta_۰$، عرض از مبدا و $\boldsymbol{\beta} = (\boldsymbol{\beta}_۱, \ldots, \boldsymbol{\beta}_q)^\top$ که $\boldsymbol{\beta}_g$ بردار پارامتر متناظر با g امین پیش‌گو است.

مایر و همکاران (۲۰۰۸) همچنین یک الگوریتم کاهش مختصات برای برازش مدل‌های لاسو گروهی پیشنهاد دادند. با این حال، آن‌ها تنها حالت خاص در گروه‌هایی که متعامد هستند را مد نظر قرار دادند. برآوردگر لاسو گروهی لجستیک $\hat{\boldsymbol{\beta}}_\lambda$ کمینه کننده تابع محدب

$$S_\lambda(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=۱}^{q} s(df_j) \left\|\boldsymbol{\beta}_g\right\|_۲ \tag{۱۶.۴}$$

است که $\ell(\cdot)$، تابع لگاریتم درستمایی به‌صورت

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \eta(\mathbf{x}_i) - \log(1 + \exp\{\eta(\mathbf{x}_i)\}) \qquad (۱۷.۴)$$

است. پارامتر منظم‌سازی $\lambda \geq 0$ اندازه جریمه را کنترل می‌کند. عرض از مبدا، جریمه نمی‌شود. کمینه در رابطه (۱۶.۴) به‌دست می‌آید. تابع $s(\cdot)$ استفاده شده است تا مقیاس جریمه را نسبت به بعد پارامتر بردار $\boldsymbol{\beta}_g$ دهد. آن‌ها $s(df_g) = df_g^{\frac{1}{2}}$ پیشنهاد کردند که تضمین می‌کند عبارت جریمه از مرتبه تعداد پارامترهای df_g است. یوآن و لین (۲۰۰۶) الگوریتمی برای دست‌یابی به پاسخ ارائه کرده‌اند.

مایر و همکاران (۲۰۰۸) لاسو گروهی را برای مدل‌های لجستیک بررسی کرده‌اند.

۴.۴ شبیه‌سازی

هدف از مطالعه این شبیه‌سازی مقایسه روش‌های انتخاب متغیر لاسو، الاستیک‌نت، لاسو گروهی و بریج گروهی است. این روش‌ها در رگرسیون لجستیک با خروجی دودویی بررسی شده است. عملکرد هر روش بر حسب دقت پیش‌گویی و تعداد متغیرهای بااهمیت یا متغیرهایی که معنی‌دار نیستند و به‌درستی تشخیص داده می‌شوند، اندازه‌گیری می‌شود. داده‌های شبیه‌سازی شده از مدل رگرسیون لجستیک

$$\text{logit}(p_i) = \log(\frac{p_i}{1 - p_i}) = \beta_0 + \mathbf{X}\boldsymbol{\beta}$$

تولید شده‌اند که در آن X_i ها پیش‌گوهای تولید شده از توزیع نرمال استاندارد با میانگین صفر هستند، Y_i ها داده‌های به‌صورت دودویی توزیع شده هستند که برای $i = 1, \cdots, n$، $Y_i \sim \text{Bin}(n_i, p_i)$، یک توزیع دوجمله‌ای با پارامترهای n_i و p_i است. مثال‌های شبیه‌سازی مبتنی بر معیارهای جدول ۱.۴ ساخته شده‌اند.

برای معیار اندازه‌های متفاوت اثر، نسبت بخت در نظر گرفته شده است. برای مثال، فرض کنید X_i یک متغیر دودویی باشد که منظور از $X_i = 1$، زن و منظور از $X_i = 0$، مرد است. می‌توان نوشت

$$\begin{aligned} \log(\text{odds}_F) = \log(\frac{p_{iF}}{1 - p_{iF}}) &= \beta_0 + \beta_1 * 1 \\ \log(\text{odds}_M) = \log(\frac{p_{iM}}{1 - p_{iM}}) &= \beta_0 + \beta_1 * 0 \end{aligned}$$

بنابراین

$$\begin{aligned} \log(\frac{\text{odds}_F}{\text{odds}_M}) &= \beta_1 \\ \text{odds ratio} = \frac{\text{odds}_F}{\text{odds}_M} &= \exp\{\beta_1\} \end{aligned}$$

جدول ۱.۴: معیارهای شبیه‌سازی برای مثال‌های مختلف

معیار شبیه‌سازی	سطوح	مقادیر عددی
اثر	بزرگ	$\log(۱/۸)$
	متوسط	$\log(۱/۶)$
	کوچک	$\log(۱/۲)$
ساختار گروهی	همه_داخل_همه_خارج	
	ن_همه_داخل_همه_خارج	
pبزرگ، n کوچک	p بزرگ‌تر از n	$p = ۴۰$ $n = ۲۰$
	p کوچک‌تر از n	$p = ۴۰$ $n = ۱۰۰$
همبستگی	بالا	۰/۸
	متوسط	۰/۶
	پایین	۰/۲

$$\beta_۱ \quad = \quad \log(\text{odds ratio})$$

در رگرسیون لجستیک، از مقدار نسبت بخت بین ۱/۱ تا ۱/۵ به‌عنوان اثر کوچک تعبیر می‌شود. مقدار ۱/۶ تا ۱/۷ به‌عنوان اثر متوسط و مقدار ۱/۸ تا ۲ به‌عنوان اثر بزرگ در نظر گرفته می‌شود. بنابراین، برای اثر کوچک، متوسط و بزرگ به‌ترتیب از $\log(۱/۲)$، $\log(۱/۶)$ و $\log(۱/۸)$ استفاده شده است.

چهار مثال در اینجا ارائه شده است: اولین مثال، در مقاله لاسو (تیبشیرانی، ۱۹۹۶) برای مقایسه عملکرد پیش‌گویی رگرسیون لاسو و ریج استفاده شده است. مثال‌های دوم و چهارم از یک مقاله درباره روش‌های انتخاب متغیر گروهی (ژی و ژنگ، ۲۰۱۰) اقتباس شده است. سومین مثال از یک مقاله در مورد الاستیک‌نت (ژو و هستی، ۲۰۰۵) گرفته شده است. سه مثال آخر، یک متغیر گروهی را ایجاد می‌کند و هدف، مقایسه لاسو با روش‌های گروهی است.

برای هر مثال، مجموعه‌های داده ۱۰۰تایی شبیه‌سازی می‌شوند. هر مجموعه داده، از یک مجموعه مدل‌ساز و یک مجموعه آزمون ساخته شده‌است. مجموعه مدل‌ساز برای انتخاب پارامتر تنظیم‌کننده ایجاد شده است. سپس مدل‌ها را به مجموعه آزمون برازش داده می‌شود و میانه توان‌های دوم خطا[۹] روی مجموعه آزمون محاسبه می‌شود. همچنین تعداد فراوانی انتخاب متغیرهای مهم و فراوانی حذف متغیرهای بی‌اهمیت ثبت شده‌است. نتایج در جداول زیر خلاصه شده‌اند.

اولین معیار برای آزمون روش‌های مختلف، پیش‌گویی است. سه اندازه مختلف برای

[9] Median squared error

سنجیدن خطای پیش‌گویی استفاده شده‌است:

۱. مقدار میانه میانگین توان‌های دوم خطا (MSE) که به‌صورت

$$\mathrm{MSE} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(\circ)})^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(\circ)})$$

تعریف می‌شود.

۲. خطاهای مدل،

$$\mathrm{ME} = \mathrm{E}(\mathrm{E}(Y|x) - \hat{\mu})^۲,$$

است که $\mathrm{E}(Y|x) = p(x) = \frac{\exp(x\beta)}{۱+\exp(x\beta)}$ و $\hat{\mu}(x) = \hat{p}(x) = \frac{\exp(x\hat{\beta})}{۱+\exp(x\hat{\beta})}$.

۳. خطای دسته‌بندی یا خطای شمارش:

$$\begin{cases} \mathrm{CE} = ۱ & \text{اگر } Y_{new} = ۱ \text{و} \hat{p} < \frac{۱}{۲} \text{یا} Y_{new} = ۰ \text{و} \hat{p} > \frac{۱}{۲} \\ \mathrm{CE} = \frac{۱}{۲} & \text{اگر } \hat{p} = \frac{۱}{۲} \\ \mathrm{CE} = ۰ & \text{غیر این صورت} \end{cases}$$

معیار دوم آزمون یک روش، درصد درستی حذف متغیرهای بی‌اهمیت و درصد درستی تشخیص متغیرهای مهم است.
چهار مثال به‌صورت زیر داده شده‌اند:

- مثال ۱. ۱۰۰ مجموعه داده شامل ۱۰۰ و ۲۰۰ مشاهده به‌ترتیب در مجموعه‌های مدل‌ساز و آزمون و با ۸ پیشگو شبیه‌سازی شده است. بردار پارامتر واقعی،

$$\boldsymbol{\beta} = (\log(۱/۸), \log(۱/۲), ۰, ۰, \log(۱/۶), ۰, ۰, ۰)^\top$$

قرار داده شده است. همبستگی زوجی میان x_i و x_j، به‌صورت $\mathrm{cor}(i,j) = ۰/۵^{(i-j)}$ فرض شده است؛ بنابراین متغیرهای پیش‌گو، دارای همبستگی اتورگرسیو مرتبه اول هستند.

- مثال ۲. ۱۰۰ مجموعه داده شامل ۱۰۰ و ۴۰۰ مشاهده به‌ترتیب در مجموعه‌های مدل‌ساز و آزمون و با ۴۰ پیشگو شبیه‌سازی می‌شود. فرض می‌شود

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۱۵}, \underbrace{\log(۱/۲), \ldots, \log(۱/۲)}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۰})^\top$$

این مثال یک گروه ۱۵ متغیره با همبستگی بالا ایجاد می‌کند. همبستگی درون گروهی، مقدار بالای ۰/۸ قرار داده می‌شود. همبستگی میان گروهی متغیرها صفر است. ۵ متغیر با ضرایب ۱/۵ نسبت به بقیه مستقل و دارای اثر کوچکی روی پاسخ هستند. ۲۰ متغیر باقی‌مانده نسبت به دیگر متغیرها مستقل هستند و هیچ اثری روی پاسخ ندارند.

- مثال ۳. ۱۰۰ مجموعه داده شامل ۱۰۰ و ۴۰۰ مشاهده به‌ترتیب در مجموعه‌های مدل‌ساز و آزمون با ۴۰ پیش‌گو شبیه‌سازی شده است. بردار پارامتر واقعی

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۲), \ldots, \log(۱/۲)}_{۵}, \underbrace{\log(۱/۲), \ldots, \log(۱/۲)}_{۵}, \underbrace{\log(۱/۲), \ldots, \log(۱/۲)}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

انتخاب شده است. این مثال، سه گروه با ۵ متغیر با همبستگی بالا در هر گروه ایجاد می‌کند. اثرهای متغیرهای گروهی در هر گروه کوچک قرار داده شده‌است. همبستگی درون گروهی، مقدار بالای ۰/۸ فرض شده است. همبستگی بین گروهی متغیرها صفر است. متغیرهای در هر گروه به‌یک اندازه برای پاسخ بااهمیت می‌باشند. همچنین ۲۵ متغیر بدون هیچ اثری بر پاسخ نیز وجود دارند.

- مثال ۲.۳. همانند مثال ۱.۳ است با این تفاوت که متغیرهای گروهی، اثر متوسطی دارند. بردار واقعی پارامتر به‌صورت

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

انتخاب شده است.

- مثال ۳.۳. همانند مثال ۱.۳ است با این تفاوت که متغیرهای گروهی دارای اثرهای بزرگی هستند. بردار واقعی پارامترها به‌صورت

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۸), \ldots, \log(۱/۸)}_{۵}, \underbrace{\log(۱/۸), \ldots, \log(۱/۸)}_{۵}, \underbrace{\log(۱/۸), \ldots, \log(۱/۸)}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

انتخاب شده‌اند.

- مثال ۴.۳. همانند مثال ۲.۳ است با این تفاوت که همبستگی گروهی، یک مقدار کم ۰/۲ فرض می‌شود. بردار واقعی پارامترها به‌صورت

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

انتخاب می‌شود.

- مثال ۵.۳. همانند مثال ۲.۳ است با این تفاوت که همبستگی گروهی، مقدار متوسط ۰/۴ در نظر گرفته می‌شود. بردار واقعی پارامترها به‌صورت

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

فرض می‌شود.

- مثال ۶.۳. همانند مثال ۲.۳ است با این تفاوت که تعداد مشاهدات برای ایجاد حالت $p > n$، $n = ۲۰$ قرار داده می‌شود. بردار واقعی پارامترها به‌صورت

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶)}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

در نظر گرفته می‌شود.

- مثال ۱.۴. ۱۰۰ مجموعه داده شامل ۱۰۰ و ۲۰۰ مشاهده در مجموعه‌های مدل‌ساز و آزمون با ۴۰ پیشگو شبیه‌سازی می‌شود. فرض کنید

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۲), \ldots, \log(۱/۲), ۰, ۰}_{۵}, \underbrace{\log(۱/۲), \ldots, \log(۱/۲), ۰, ۰}_{۵},$$
$$\underbrace{\log(۱/۲), \ldots, \log(۱/۲)۰, ۰}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

این مثال نیز، سه گروه متغیرهای با همبستگی بالا ایجاد می‌کند. اثرهای متغیرهای گروهی در هر گروه، کوچک در نظر گرفته می‌شود. همبستگی درون گروهی، یک مقدار بالا (۰/۸) قرار داده می‌شود. با این حال در هر گروه، سه متغیر بااهمیت و دو متغیر با اثر صفر وجود دارد. ۲۵ متغیر باقی‌مانده، متغیرهای مستقل با اثر صفر هستند.

- مثال ۲.۴. همانند مثال ۱.۴ است به‌جز با این تفاوت که متغیرهای گروهی، دارای اثرهای متوسط هستند. فرض کنید

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۶), \ldots, \log(۱/۶), ۰, ۰}_{۵}, \underbrace{\log(۱/۶), \ldots, \log(۱/۶), ۰, ۰}_{۵},$$
$$\underbrace{\log(۱/۶), \ldots, \log(۱/۶)۰, ۰}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

- مثال ۳.۴. همانند مثال ۱.۴ است با این تفاوت که متغیرهای گروهی، دارای اثرهای بزرگ هستند. فرض کنید

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۸), \ldots, \log(۱/۸), ۰, ۰}_{۵}, \underbrace{\log(۱/۸), \ldots, \log(۱/۸), ۰, ۰}_{۵},$$
$$\underbrace{\log(۱/۸), \ldots, \log(۱/۸)۰, ۰}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

- مثال ۴.۴. همانند مثال ۱.۴ است با این تفاوت که فقط دو متغیر دارای اثر بزرگ هستند و بقیه متغیرها اثر صفر دارند. فرض کنید

$$\boldsymbol{\beta} = (\underbrace{\log(۱/۸), \log(۱/۸), ۰, ۰, ۰}_{۵}, \underbrace{\log(۱/۸), \log(۱/۸), ۰, ۰, ۰}_{۵},$$
$$\underbrace{\log(۱/۸), \log(۱/۸)۰, ۰, ۰}_{۵}, \underbrace{۰, \ldots, ۰}_{۲۵})^\top$$

جدول‌های ۲.۴ و ۳.۴ میانه (میانگین) توان دوم خطا برای ۱۰۰ شبیه‌سازی را نشان می‌دهد. ملاحظه می‌شود که الاستیک‌نت در همه مثال‌ها، MSE کمتری نسبت به لاسو دارد. بنابراین الاستیک‌نت در برآورد ضرایب با وجود هم‌خطی، از لاسو دقیق‌تر است. روش‌های انتخاب متغیر لاسو گروهی در همه مثال‌ها به‌طور قابل ملاحظه‌ای بهتر از لاسو و الاستیک‌نت عمل می‌کند. بریج گروهی در بعضی مثال‌ها (مثال ۱.۲ و ۱.۳ و ۱.۴) MSE خیلی بزرگی دارد، بنابراین روش خیلی خوبی نیست. لاسو گروهی، در برآورد ضرایب برای انتخاب متغیر گروهی دقیق‌ترین روش است. عملکرد آن درطول همه مثال‌ها بسیار پایدار است. در مثال‌های ۱.۱، ۱.۲، ۱.۳، و ۱.۴ مقدار کاهش لاسو گروهی در مقایسه MSE با لاسو ٪۶۲/۸۲، ٪۵۸/۳۲، ٪۶۷/۷۸ و ٪۲۴/۱۲ است. جدول‌های ۴.۴ و ۵.۴، خطای مدل (ME) برای ۱۰۰ شبیه‌سازی را نشان می‌دهد. لاسو گروهی هنوز بهتر عمل می‌کند و در بیشتر مثال‌ها کمترین ME را دارد. جدول‌های ۶.۴ و ۷.۴، خطای طبقه‌بندی اشتباه [10] (MCE) را نشان می‌دهد. برآوردگر لاسو گروهی، کمترین MCE را در مقایسه با سایر روش‌ها دارد.

جدول‌های ۸.۴ و ۹.۴، تعداد متغیرهای مهم به‌درستی تشخیص داده شده و تعداد متغیرهای بی‌اهمیت به‌درستی حذف شده را نشان می‌دهد. لاسو مدل‌های خیلی تُنُکی تولید می‌کند اما برای همخطی کار نمی‌کند. الاستیک‌نت، لاسو را وقتی پیشگوها همبسته هستند، بهبود می‌دهد. اما الاستیک‌نت، پنج تا از اثرهای مهم با ضرایب کوچک $\log(۱/۲)$ در مثال ۲.۱ را از دست داده است. در مثال ۲.۱ لاسو گروهی به‌لحاظ تعداد متغیرهای به‌درستی تشخیص داده شده، الاستیک‌نت را بهبود می‌دهد. مثال ۳.۱ موردی است که در آن همه متغیرهای در یک گروه مهم هستند. که این یک طرح «همه داخل، همه خارج» را نشان می‌دهد، به‌طور مثال وقتی یک متغیر در یک گروه انتخاب شده باشد، همه متغیرهای دیگر در همان گروه هم باید انتخاب شوند. لاسو گروهی همه متغیرهای مهم را انتخاب کرده است، درحالی که لاسو و بریج گروهی، بعضی از متغیرهای در گروه را از دست داده‌اند. بنابراین می‌توان مشاهده کرد که لاسو گروهی، گروه کامل را تحمیل می‌کند. در مثال ۳، چگونگی اثر گذاری اثرهای کوچک، متوسط و بزرگ را بر انتخاب متغیر گروهی مقایسه شده است. در مثال ۳.۳ که پنج اثر کوچک درون یک گروه وجود دارند، لاسو گروهی با موفقیت همه اثرهای کوچک را انتخاب می‌کند درحالی که لاسو فقط دو تا را انتخاب می‌کند، و الاستیک‌نت و بریج گروهی، سه تا انتخاب می‌کنند. از این رو، لاسو گروهی بهترین روش برای گروه‌بندی کامل است. برای اثرهای کوچک درون یک گروه مشترک با یک اثر بزرگ، به‌خوبی کار می‌کند.

مثال ۴ یک موقعیت تُنُک ایجاد می‌کند که در آن ۵ متغیر در یک گروه وجود دارند و ۳تای آن‌ها اثر مهم دارند. این نشان‌دهنده یک طرح «نه_همه_داخل_همه_خارج» است. لاسو گروهی همه متغیرها از هر گروه را انتخاب می‌کند حتی بعضی از آن‌ها اثرهای صفر دارند. بریج گروهی ۴ متغیر از هر گروه انتخاب می‌کند. همچنین در مثال ۴.۴، اگر چه ۵ متغیر در یک گروه وجود دارند و ۲ تا از آن‌ها اثر بزرگ با $\log(۲)$ هستند، لاسو گروهی همه ۵ متغیر

[10] Misclassification error (MCE)

را انتخاب می‌کند. این روش، متغیرها را نه بر اساس سطح مجزا بلکه براساس سطح گروهی انتخاب می‌کند.

وقتی روش جریمه‌شده برای گروه به‌کار گرفته می‌شود، ساختار داده باید در نظر گرفته شود. اگر محقق انتظار دارد که تناسب اعضای غیرصفر گروه بزرگتر از ۵٪ باشد، لاسو گروهی استفاده می‌شود. اگر محقق، انتظار دارد که این تناسب نزدیک به ۵٪ باشد، ممکن است بخواهد از بریج گروهی استفاده کند.

این روش‌ها، با همبستگی گروهی و اندازه نمونه مختلف، آزمون می‌شوند. از مثال ۴.۳ و ۵.۳ که در آن‌ها همبستگی‌ها متوسط و کوچک قرار داده می‌شود، تفاوت بزرگی برای روش‌های مختلف از نظر دقت پیشگویی و تعداد متغیرهای به‌درستی انتخاب شده وجود ندارد، پیدا می‌شود. با این حال اگر اندازه نمونه کوچک همانند مثال ۶.۳ قرار داده شود، دقت پیش‌گویی به‌طور قابل توجهی برای هر روش کاهش می‌یابد. به‌لحاظ تعداد متغیرهای به‌درستی انتخاب شده، مشاهده می‌شود که موضوع همخطی برای لاسو واضح‌تر است. لاسو گروهی تنها یک متغیر مهم از گروه انتخاب می‌کند درحالی که لاسو گروهی ۵ تا انتخاب می‌کند. بنابراین، لاسو گروهی هنوز از بقیه مدل‌های در طرح «همه داخل، همه خارج»، p بزرگ و n کوچک بهتر عمل می‌کند.

جدول ۲.۴: MSE: میانه، میانگین، چارک‌های اول و سوم میانگین توان دوم خطا برای ۱۰۰ شبیه‌سازی

مثال‌ها	روش‌ها	میانه	میانگین	$Q1(25\%)$	$Q3(75\%)$
۱.۱	لاسو	۰/۳۹	۰/۴۱	۰/۳۲	۰/۵۴
	الاستیک‌نت	۰/۳۸	۰/۴۰	۰/۳۰	۰/۵۲
	بریج گروهی	۰/۱۲	۰/۱۹	۰/۰۷	۰/۲۷
	لاسو گروهی	۰/۱۵	۰/۱۷	۰/۰۹	۰/۲۵
۱.۲	لاسو	۱/۹۶	۱/۹۴	۱/۷۵	۲۰/۹
	الاستیک‌نت	۱/۴۸	۱/۴۹	۱/۲۶	۱/۶۸
	بریج گروهی	۱۲/۸۲	۲۸/۸۵	۸/۱۸	۲۵/۳۲
	لاسو گروهی	۰/۸۲	۱/۱۴	۰/۶۹	۱/۰۶
۱.۳	لاسو	۰/۳۸	۰/۳۷	۰/۳۳	۰/۴۲
	الاستیک‌نت	۰/۳۱	۰/۳۲	۰/۲۷	۰/۳۵
	بریج گروهی	۰/۴۰	۰/۶۹	۰/۳۲	۰/۷۲
	لاسو گروهی	۰/۱۱	۰/۱۳	۰/۰۹	۰/۱۴
۲.۳	لاسو	۱/۷۳	۱/۷۵	۱/۵۴	۱/۹۹
	الاستیک‌نت	۱/۴۴	۱/۴۵	۱/۲۰	۱/۶۴
	بریج گروهی	۳/۹۹	۴/۳۸	۲/۶۶	۵/۳۷
	لاسو گروهی	۰/۶۷	۰/۸۵	۰/۵۶	۰/۸۵
۳.۳	لاسو	۲/۶۵	۲/۶۴	۲/۲۹	۲/۹۶
	الاستیک‌نت	۲/۳۵	۲/۲۶	۱/۹۷	۲/۶۱
	بریج گروهی	۵/۹۹	۶/۵۹	۳/۹۰	۷/۴۴
	لاسو گروهی	۱/۲۳	۱/۳۷	۰/۹۳	۱/۵۸
۴.۳	لاسو	۲/۲۲	۲/۲۵	۱/۹۵	۲/۵۸
	الاستیک‌نت	۲/۱۰	۲/۰۶	۱/۷۵	۲/۳۳
	بریج گروهی	۱/۳۳	۱/۴۴	۱/۰۸	۱/۷۲
	لاسو گروهی	۰/۸۳	۰/۸۲	۰/۵۹	۰/۹۹

جدول ۳.۴: MSE: میانه، میانگین، چارک‌های اول و سوم میانگین توان دوم خطا برای ۱۰۰ شبیه‌سازی (ادامه...)

مثال‌ها	روش‌ها	میانه	میانگین	Q۱(۲۵٪)	Q۳(۷۵٪)
۵.۳	لاسو	۱/۸۹	۱/۹۲	۱/۶۸	۲/۲۱
	الاستیک‌نت	۱/۷۱	۱/۷۳	۱/۴۵	۱/۹۸
	بریج گروهی	۱/۶۴	۱/۸۳	۱/۲۷	۲/۲۴
	لاسو گروهی	۰/۷۲	۰/۷۸	۰/۵۸	۰/۸۹
۶.۳	لاسو	۳/۱۸	۳/۹۰	۳/۰۵	۳/۳۱
	الاستیک‌نت	۳/۰۴	۳/۱۰	۲/۷۶	۳/۲۶
	بریج گروهی	۷/۴۸	۹/۸۵	۵/۳۸	۱۱/۵۸
	لاسو گروهی	۳/۵۵	۴/۷۲	۲/۵۲	۶/۲۳
۱.۴	لاسو	۰/۲۹	۰/۲۸	۰/۲۷	۰/۳۰
	الاستیک‌نت	۰/۲۸	۰/۲۷	۰/۲۷	۰/۳۰
	بریج گروهی	۰/۳۱	۰/۴۰	۰/۳۰	۰/۳۷
	لاسو گروهی	۰/۲۲	۰/۲۳	۰/۱۸	۰/۲۵
۲.۴	لاسو	۱/۲۵	۱/۲۷	۱/۱۲	۱/۴۴
	الاستیک‌نت	۱/۱۳	۱/۱۵	۱/۰۲	۱/۳۱
	بریج گروهی	۶/۳۸	۸/۷۰	۱/۴۵	۱۰/۸۱
	لاسو گروهی	۰/۹۵	۱/۴۸	۰/۸۴	۱/۰۹
۳.۴	لاسو	۱/۷۸	۱/۸۵	۱/۵۹	۲/۱۰
	الاستیک‌نت	۱/۷۲	۱/۷۳	۱/۵۴	۱/۹۴
	بریج گروهی	۱۳/۰۰	۱۷/۵۳	۷/۳۴	۱۹/۱۱
	لاسو گروهی	۱/۳۷	۱/۷۵	۱/۲۶	۱/۵۰
۴.۴	لاسو	۱/۳۳	۱/۳۷	۱/۱۹	۱/۵۷
	الاستیک‌نت	۱/۳۳	۱/۳۳	۱/۱۸	۱/۴۸
	بریج گروهی	۱/۷۳	۶/۱۸	۱/۰۳	۷/۴۹
	لاسو گروهی	۱/۲۵	۱/۳۲	۱/۱۶	۱/۳۷

جدول ۴.۴: ME: میانه، میانگین، چارک‌های اول و سوم خطا مدل برای ۱۰۰ شبیه‌سازی

مثال‌ها	روش‌ها	میانه	میانگین	$Q1(25\%)$	$Q3(75\%)$
۱.۱	لاسو	۴/۱۷	۴/۴۴	۳/۱۷	۶/۰۰
	الاستیک‌نت	۴/۰۱	۴/۲۱	۲/۸۳	۵/۸۳
	بریج گروهی	۱/۰۳	۱/۵۱	۰/۵۵	۲/۲۸
	لاسو گروهی	۱/۱۱	۱/۳۴	۰/۶۶	۱/۸۸
۱.۲	لاسو	۷/۸۰	۷/۸۲	۵/۰۷	۱۰/۳۲
	الاستیک‌نت	۶/۲۶	۶/۵۵	۴/۲۶	۸/۲۹
	بریج گروهی	۷/۶۴	۷/۸۵	۶/۶۵	۸/۸۲
	لاسو گروهی	۲/۴۷	۲/۷۴	۱/۹۰	۳/۴۲
۱.۳	لاسو	۱۳/۴۵	۱۵/۲۰	۱۰/۵۹	۱۹/۴۴
	الاستیک‌نت	۱۲/۷۵	۱۳/۵۷	۱۰/۰۷	۱۵/۷۸
	بریج گروهی	۷/۲۹	۹/۴۰	۵/۲۵	۱۱/۹۲
	لاسو گروهی	۳/۳۹	۳/۷۴	۲/۲۶	۴/۵۰
۲.۳	لاسو	۱۱/۸۲	۱۲/۴۸	۹/۰۸	۱۵/۸۸
	الاستیک‌نت	۱۰/۰۳	۱۰/۶۷	۷/۱۶	۱۳/۱۹
	بریج گروهی	۷/۲۱	۷/۳۷	۶/۰۹	۸/۴۷
	لاسو گروهی	۳/۴۲	۳/۷۱	۲/۴۶	۴/۶۰
۳.۳	لاسو	۱۰/۸۹	۱۱/۶۹	۸/۶۵	۱۴/۳۷
	الاستیک‌نت	۱۰/۰۴	۹/۹۳	۷/۳۱	۱۲/۴۲
	بریج گروهی	۷/۱۱	۷/۳۲	۶/۰۰	۸/۴۲
	لاسو گروهی	۳/۷۵	۴/۰۹	۲/۶۸	۵/۲۱
۴.۳	لاسو	۲۳/۱۶	۲۴/۱۹	۱۸/۷۲	۲۹/۰۸
	الاستیک‌نت	۲۰/۵۷	۲۰/۹۲	۱۵/۷۲	۲۴/۶۵
	بریج گروهی	۸/۱۷	۹/۵۱	۶/۵۸	۹/۹۴
	لاسو گروهی	۵/۶۸	۵/۹۳	۴/۰۳	۷/۰۹

جدول ۵.۴: ME: میانه، میانگین، چارک‌های اول و سوم خطا مدل برای ۱۰۰ شبیه‌سازی (ادامه...)

مثال‌ها	روش‌ها	میانه	میانگین	$Q1(25\%)$	$Q3(75\%)$
۵.۳	لاسو	۱۷/۹۲	۱۸/۷۵	۱۴/۲۸	۲۲/۷۵
	الاستیک‌نت	۱۴/۶۴	۱۵/۸۸	۱۱/۳۵	۱۸/۸۳
	بریج گروهی	۷/۶۱	۸/۴۴	۶/۵۲	۸/۸۷
	لاسو گروهی	۴/۳۲	۴/۸۱	۳/۴۷	۵/۷۷
۶.۳	لاسو	۲/۱۱	۲/۰۶	۱/۳۲	۲/۷۰
	الاستیک‌نت	۱/۹۹	۱/۹۶	۱/۳۵	۲/۵۹
	بریج گروهی	۱/۳۴	۱/۵۰	۰/۹۸	۱/۹۱
	لاسو گروهی	۱/۹۲	۱/۲۴	۰/۷۹	۱/۵۸
۱.۴	لاسو	۶/۲۳	۵/۸۳	۵/۰۶	۶/۸۹
	الاستیک‌نت	۵/۹۹	۵/۷۳	۴/۹۱	۶/۸۹
	بریج گروهی	۵/۷۵	۵/۶۱	۴/۱۹	۶/۶۹
	لاسو گروهی	۲/۶۱	۲/۸۷	۱/۹۲	۳/۸۶
۲.۴	لاسو	۶/۷۵	۶/۷۹	۴/۹۵	۸/۶۷
	الاستیک‌نت	۵/۹۹	۶/۱۰	۴/۴۷	۷/۷۸
	بریج گروهی	۶/۸۷	۶/۴۴	۴/۲۸	۸/۱۱
	لاسو گروهی	۲/۷۱	۳/۰۸	۲/۰۶	۳/۵۸
۳.۴	لاسو	۶/۴۰	۶/۷۰	۵/۰۶	۸/۲۳
	الاستیک‌نت	۵/۷۹	۶/۰۹	۴/۵۶	۷/۲۴
	بریج گروهی	۷/۲۸	۷/۲۹	۶/۰۲	۸/۸۹
	لاسو گروهی	۲/۵۸	۲/۸۷	۲/۰۲	۳/۲۲
۴.۴	لاسو	۷/۰۱	۷/۲۴	۴/۹۶	۸/۸۲
	الاستیک‌نت	۶/۳۹	۶/۶۷	۵/۰۹	۷/۸۱
	بریج گروهی	۶/۷۰	۵/۹۵	۲/۷۷	۸/۴۹
	لاسو گروهی	۳/۱۱	۳/۳۸	۲/۶۶	۳/۹۶

جدول ۶.۴: MCE: میانه، میانگین، چارک‌های اول و سوم خطای طبقه‌بندی اشتباه برای ۱۰۰ شبیه‌سازی

مثال‌ها	روش‌ها	میانه	میانگین	$Q1(25\%)$	$Q3(75\%)$
۱.۱	لاسو	۰/۱۹۰۰	۰/۱۸۵۶	۰/۱۵۰۰	۰/۲۲۱۲۵
	الاستیک‌نت	۰/۱۸۰۰	۰/۱۷۸۷	۰/۱۵۰۰	۰/۲۱۲۵
	بریج گروهی	۰/۱۷۰۰	۰/۱۷۵۶	۰/۱۴۷۵	۰/۲۱۰۰
	لاسو گروهی	۰/۱۷۰۰	۰/۱۶۸۶	۰/۱۴۰۰	۰/۲۰۰۰
۱.۲	لاسو	۰/۰۴۰۰	۰/۰۴۰۳	۰/۰۳۵۰	۰/۰۴۷۵
	الاستیک‌نت	۰/۰۴۰۰	۰/۰۳۹۷	۰/۰۳۲۵	۰/۰۴۷۵
	بریج گروهی	۰/۰۳۰۰	۰/۰۲۸۵	۰/۰۲۰۰	۰/۰۳۷۵
	لاسو گروهی	۰/۰۳۷۵	۰/۰۳۷۴	۰/۰۳۰۰	۰/۰۴۵۰
۱.۳	لاسو	۰/۱۳۷۵	۰/۱۴۳۰	۰/۱۲۶۹	۰/۱۵۷۵
	الاستیک‌نت	۰/۱۳۲۵	۰/۱۳۷۴	۰/۱۲۲۵	۰/۱۴۷۵
	بریج گروهی	۰/۱۳۲۵	۰/۱۴۰۱	۰/۱۲۰۰	۰/۱۵۲۵
	لاسو گروهی	۰/۱۳۱۳	۰/۱۳۰۰	۰/۱۱۶۹	۰/۱۴۲۵
۲.۳	لاسو	۰/۰۶۵۰	۰/۰۶۷۰	۰/۰۵۷۵	۰/۰۷۲۵
	الاستیک‌نت	۰/۰۶۵۰	۰/۰۶۶۰	۰/۰۵۲۵	۰/۰۷۲۵
	بریج گروهی	۰/۰۵۵۰	۰/۰۵۶۱	۰/۰۴۵۰	۰/۰۶۵۶
	لاسو گروهی	۰/۰۶۳۸	۰/۶۳۳۳	۰/۰۵۲۵	۰/۰۷۲۵
۳.۳	لاسو	۰/۰۵۵۰	۰/۰۵۴۱	۰/۰۴۶۹	۰/۰۶۲۵
	الاستیک‌نت	۰/۰۵۵۰	۰/۰۵۳۲	۰/۰۴۶۹	۰/۰۶۰۰
	بریج گروهی	۰/۰۴۳۸	۰/۰۴۴۲	۰/۰۳۵۰	۰/۰۵۵۰
	لاسو گروهی	۰/۰۵۰۰	۰/۰۵۰۷	۰/۰۴۵۰	۰/۰۵۷۵
۴.۳	لاسو	۰/۱۱۲۵	۰/۱۱۴۴	۰/۰۹۵۰	۰/۱۲۷۵
	الاستیک‌نت	۰/۱۰۲۵	۰/۱۰۳۱	۰/۰۹۰۰	۰/۱۱۵۰
	بریج گروهی	۰/۰۸۷۵	۰/۰۸۹۱	۰/۰۷۵۰	۰/۱۰۲۵
	لاسو گروهی	۰/۰۹۰۰	۰/۰۸۹۰	۰/۰۷۹۴	۰/۱۰۰۰

جدول ۷.۴: MCE: میانه، میانگین، چارک‌های اول و سوم خطای طبقه‌بندی اشتباه برای ۱۰۰ شبیه‌سازی (ادامه...)

مثال‌ها	روش‌ها	میانه	میانگین	Q۱(۲۵٪)	Q۳(۷۵٪)
۵.۳	لاسو	۰/۰۸۷۵	۰/۸۸۹۸	۰/۰۷۶۹	۰/۱۰۰۰
	الاستیک‌نت	۰/۰۸۳۸	۰/۰۸۳۲	۰/۰۷۲۵	۰/۰۹۵۰
	بریج گروهی	۰/۰۷۵۰	۰/۰۷۵۱	۰/۰۶۰۰	۰/۰۸۷۵
	لاسو گروهی	۰/۰۷۷۵	۰/۰۷۷۱	۰/۰۶۵۰	۰/۰۸۸۱
۶.۳	لاسو	۰/۰۵۰۰	۰/۰۷۷۰	۰/۰۰۰۰	۰/۱۵۰۰
	الاستیک‌نت	۰/۰۵۰۰	۰/۰۵۹۵	۰/۰۰۰۰	۰/۱۰۰۰
	بریج گروهی	۰/۰۰۰۰	۰/۰۰۲۰	۰/۰۰۰۰	۰/۰۰۰۰
	لاسو گروهی	۰/۰۰۰۰	۰/۰۰۷۰	۰/۰۰۰۰	۰/۰۰۰۰
۱.۴	لاسو	۰/۱۹۵۰	۰/۱۹۷۳	۰/۱۶۵۰	۰/۲۳۲۵
	الاستیک‌نت	۰/۱۸۰۰	۰/۱۹۴۳	۰/۱۶۵۰	۰/۲۲۶۳
	بریج گروهی	۰/۱۹۵۰	۰/۱۹۸۸	۰/۱۶۵۰	۰/۲۴۰۰
	لاسو گروهی	۰/۱۶۰۰	۰/۱۶۲۳	۰/۱۴۰۰	۰/۱۸۵۰
۲.۴	لاسو	۰/۰۹۵۰	۰/۰۹۷۶	۰/۰۸۰۰	۰/۱۱۱۳
	الاستیک‌نت	۰/۰۹۵۰	۰/۰۹۵۰	۰/۰۸۰۰	۰/۱۱۰۰
	بریج گروهی	۰/۰۸۰۰	۰/۰۷۷۲	۰/۰۵۵۰	۰/۰۹۵۰
	لاسو گروهی	۰/۰۸۷۵	۰/۰۸۹۴	۰/۰۷۵۰	۰/۱۰۱۳
۳.۴	لاسو	۰/۰۸۰۰	۰/۰۷۹۳	۰/۰۶۵۰	۰/۰۹۰۰
	الاستیک‌نت	۰/۰۸۰۰	۰/۰۷۹۲	۰/۰۶۵۰	۰/۰۹۰۰
	بریج گروهی	۰/۰۶۰۰	۰/۰۵۹۳	۰/۰۴۰۰	۰/۰۷۵۰
	لاسو گروهی	۰/۰۷۰۰	۰/۰۷۳۶	۰/۰۶۰۰	۰/۰۸۵۰
۴.۴	لاسو	۰/۱۱۰۰	۰/۱۱۱۶	۰/۰۹۵۰	۰/۱۳۱۳
	الاستیک‌نت	۰/۰۱۰۴	۰/۱۰۸۶	۰/۰۹۰۰	۰/۱۲۶۳
	بریج گروهی	۰/۰۸۵۰	۰/۰۹۳۵	۰/۰۷۵۰	۰/۱۱۶۳
	لاسو گروهی	۰/۱۰۵۰	۰/۱۰۲۳	۰/۰۸۵۰	۰/۱۱۶۳

جدول ۸.۴: درصد به‌درستی تشخیص دادن متغیرهای مهم و به‌درستی حذف شدن متغیرهای بی‌اهمیت

مثال‌ها	روش‌ها	درصد متغیرهای غیرصفر	درصد متغیرهای صفر
۱.۱	لاسو	۲(۶۶/۷٪)	۵(۱۰۰٪)
	الاستیک‌نت	۲(۶۶/۷٪)	۵(۱۰۰٪)
	بریج گروهی	۲(۶۶/۷٪)	۵(۱۰۰٪)
	لاسو گروهی	۲(۶۶/۷٪)	۴(۸۰٪)
۱.۲	لاسو	۱۱,۰(۵۵٪)	۲۰(۱۰۰٪)
	الاستیک‌نت	۱۵,۰(۷۵٪)	۲۰(۱۰۰٪)
	بریج گروهی	۱۵,۴(۹۵٪)	۴(۲۰٪)
	لاسو گروهی	۱۵,۲(۸۵٪)	۱۶(۸۰٪)
۱.۳	لاسو	۲,۲,۲(۴۰٪)	۲۵(۱۰۰٪)
	الاستیک‌نت	۴,۴,۴(۸۰٪)	۲۵(۱۰۰٪)
	بریج گروهی	۳,۳,۳(۶۰٪)	۲۵(۱۰۰٪)
	لاسو گروهی	۵,۵,۵(۱۰۰٪)	۲۴(۹۶٪)
۲.۳	لاسو	۴,۴,۴(۸۰٪)	۲۵(۱۰۰٪)
	الاستیک‌نت	۵,۵,۵(۱۰۰٪)	۲۵(۱۰۰٪)
	بریج گروهی	۵,۵,۵(۱۰۰٪)	۶(۲۴٪)
	لاسو گروهی	۵,۵,۵(۱۰۰٪)	۲۲(۸۸٪)
۳.۳	لاسو	۴,۴,۴(۸۰٪)	۲۵(۱۰۰٪)
	الاستیک‌نت	۵,۵,۵(۱۰۰٪)	۲۵(۱۰۰٪)
	بریج گروهی	۵,۵,۵(۱۰۰٪)	۷(۲۸٪)
	لاسو گروهی	۵,۵,۵(۱۰۰٪)	۲۲(۸۸٪)
۴.۳	لاسو	۴,۴,۴(۸۰٪)	۲۵(۱۰۰٪)
	الاستیک‌نت	۵,۵,۵(۱۰۰٪)	۲۵(۱۰۰٪)
	بریج گروهی	۵,۵,۵(۱۰۰٪)	۵(۲۰٪)
	لاسو گروهی	۵,۵,۵(۱۰۰٪)	۲۲(۸۸٪)

جدول ۹.۴: درصد به‌درستی تشخیص دادن متغیرهای مهم و به‌درستی حذف شدن متغیرهای بی‌اهمیت (ادامه...)

مثال‌ها	روش‌ها	درصد متغیرهای غیرصفر	درصد متغیرهای صفر
۵.۳	لاسو	۴,۵,۴(۸۷٪)	۲۵(۱۰۰٪)
	الاستیک‌نت	۵,۵,۵(۱۰۰٪)	۲۵(۱۰۰٪)
	بریج گروهی	۵,۵,۵(۱۰۰٪)	۴(۱۰۰٪)
	لاسو گروهی	۵,۵,۵(۱۰۰٪)	۲۱(۸۴٪)
۶.۳	لاسو	۱,۱,۱(۲۰٪)	۲۵(۱۰۰٪)
	الاستیک‌نت	۱,۲,۱(۲۷٪)	۲۴(۹۶٪)
	بریج گروهی	۳,۳,۳(۶۰٪)	۱۹(۷۶٪)
	لاسو گروهی	۵,۵,۵(۱۰۰٪)	۱۸(۷۲٪)
۱.۴	لاسو	۰,۰,۰(۰٪)	۲,۲,۲ + ۲۵ = ۳۱(۱۰۰٪)
	الاستیک‌نت	۱,۰,۰(۰٪)	۲,۲,۲ + ۲۵ = ۳۱(۱۰۰٪)
	بریج گروهی	۰,۰,۰(۰٪)	۲,۲,۲ + ۲۵ = ۳۱(۱۰۰٪)
	لاسو گروهی	۳,۳,۳(۱۰۰٪)	۰,۰,۰ + ۲۱ = ۲۱(۶۸٪)
۲.۴	لاسو	۲,۲,۲(۶۷٪)	۲,۱,۲ + ۲۵ = ۳۰(۹۷٪)
	الاستیک‌نت	۳,۳,۳(۱۰۰٪)	۱,۱,۱ + ۲۵ = ۲۸(۹۰٪)
	بریج گروهی	۳,۳,۳(۱۰۰٪)	۰,۰,۰ + ۶ = ۶(۲۰٪)
	لاسو گروهی	۳,۳,۳(۱۰۰٪)	۰,۰,۰ + ۱۹ = ۱۹(۶۱٪)
۳.۴	لاسو	۲,۲,۲(۶۸٪)	۲,۱,۲ + ۲۵ = ۳۰(۹۸٪)
	الاستیک‌نت	۳,۳,۳(۱۰۰٪)	۱,۱,۱ + ۲۵ = ۲۸(۹۰٪)
	بریج گروهی	۳,۳,۳(۱۰۰٪)	۰,۰,۰ + ۶ = ۶(۲۰٪)
	لاسو گروهی	۳,۳,۳(۱۰۰٪)	۰,۰,۰ + ۱۹ = ۱۹(۶۱٪)
۴.۴	لاسو	۲,۱,۲(۸۳/۳۳٪)	۳,۲,۳ + ۲۵ = ۳۳(۹۷٪)
	الاستیک‌نت	۲,۲,۲(۱۰۰٪)	۱,۲,۲ + ۲۵ = ۳۰(۸۸٪)
	بریج گروهی	۲,۲,۲(۱۰۰٪)	۱,۱,۱ + ۲۴ = ۲۷(۷۹٪)
	لاسو گروهی	۲,۲,۲(۱۰۰٪)	۰,۰,۰ + ۲۱ = ۲۱(۶۲٪)

جدول ۱۰.۴: تعداد متغیرهای انتخاب شده در هر گروه در مثال‌های مختلف

مثال‌ها	روش‌ها	تعداد متغیرهای انتخاب شده هر گروه		
۱.۴	لاسو	۰	۰,	۰,
	الاستیک‌نت	۱	۰,	۰,
	بریج گروهی	۰	۰,	۰,
	لاسو گروهی	۵	۵,	۵,
۲.۴	لاسو	۲	۳,	۲,
	الاستیک‌نت	۴	۴,	۴,
	بریج گروهی	۵	۵,	۵,
	لاسو گروهی	۵	۵,	۵,
۳.۴	لاسو	۲	۳,	۲,
	الاستیک‌نت	۴	۴,	۴,
	بریج گروهی	۵	۵,	۵,
	لاسو گروهی	۵	۵,	۵,
۴.۴	لاسو	۲	۲,	۲,
	الاستیک‌نت	۴	۳,	۳,
	بریج گروهی	۴	۴,	۴,
	لاسو گروهی	۵	۵,	۵,

۵.۴ مثال واقعی

۱.۵.۴ کاربردی برای آنالیز داده‌های SNP

روش‌هایی که تاکنون معرفی شده است، در عمل برای داده بعد بالا، برای مثال، مطالعات ژنومی، کاربردی هستند. مطالعات ژنومی، روشی برای تشخیص ژن‌ها در بیماری‌های انسانی هستند. این روش، ژنوم را برای تغییرات کوچک جستجو می‌کند که چندریختی‌های نوکلئوتید منفرد[۱۱] یا SNPها نامیده می‌شوند، که غالبا در افراد با بیماری خاص بیشتر از افراد بدون بیماری اتفاق می‌افتد. هر مطالعه می‌تواند در یک زمان به صدها یا هزاران SNP نگاه کند. محققان از داده‌هایی از این نوع مطالعه، برای تعیین دقیق ژن‌هایی که ممکن است خطر فردی توسعه یک بیماری حتمی شرکت کنند، استفاده می‌کنند.

[11] Single nucleotide polymorphisms

از آن‌جایی که مطالعات مرتبط با ژنوم، SNPها را در سراسر ژنوم مورد بررسی قرار می‌دهد، یک روش امیدوار کننده برای مطالعه بیماری‌های رایج و پیچیده را ارائه می‌دهند که در آن تغییرات ژنتیکی زیادی برای خطر فردی شرکت دارند. این روش، هم‌اکنون ، SNPهای مرتبط با بعضی بیماری‌ها مانند دیابت‌ها، نابهنجاری‌های قلبی و بیماری پارکینسون را مشخص کرده است. محققان امیدوارند که مطالعات ژنومی، SNPهای بیشتری مرتبط با بیماری مزمن، همچنین تغییراتی که بر شخص پاسخ دهنده به داروهای حتمی اثر می‌گذارد و بر فعل و انفعال بین ژن‌های شخص و محیط اثر می‌گذارد، شناسایی کنند.

مثالی که اینجا برای ارزیابی روش‌ها استفاده شده است، شامل داده‌های تغییر ژنتیکی SNPs از یک مطالعه کنترل موردی از بیماری ویروسی نیل غربی با ۱۷۷ مورد[۱۲] و ۲۶۲ کنترل است. یافتن همه ارتباط‌های ممکن بین SNPهای قطعی و بیماری ویروسی نیل غربی، هدف مورد بررسی است. داده‌ها در مرکز نابراسکا با همه اشخاص سفید پوست قفقازی جمع‌آوری شده‌اند. آن‌جا ۴۳۹ شخص مشارکت‌کننده در مطالعه و ۵۰۰ SNP علامت‌گذار ژنوتیپ برای اشخاص وجود دارند.

هر SNP به‌صورت صفر، ۱ و ۲ به‌ترتیب برای ژنوتیپ‌های هموزیگوت (AA)، هتروزیگوت (Aa) و جهش (نادر) هموزیگوت (aa) مطابق با فراوانی ژنوتیپ کدگذاری می‌شود. ابتدا داده‌ها برای حذف کردن SNPها که دارای نرخ تماس کمتر از ٪۹۵ هستند (نرخ ازدست رفتگی بیشتر از ٪۵) یا فراوانی آلل جزئی کمتر از ٪۵ ، مورد بررسی قرار داده می‌شوند. تعداد SNP علامت‌گذار باقی‌مانده بعد از بررسی تا ۴۰۷ کاهش یافته است. متغیر پاسخ خروجی دودویی کنترل موردی وضعیت عفونی ویروس نیل غربی است. پیشگوها ۴۰۷ SNP علامت‌گذار هستند. ابتدا آزمون عدم‌همبستگی پیوند برای بررسی وابستگی ساختاری SNPهای علامت‌گذار بین یکدیگر، اجرا می‌شود. از آن‌جایی که نتایج آزمون نشان می‌دهد که همبستگی بین مکان‌ها به‌طور نسبی پایین (۰٫۱ – ۰٫۰۱) است، روش لاسو برای SNPهای علامت‌گذار به‌طور نسبی مستقل به‌جای گروه‌بندی کردن آن‌ها و اعمال روش‌های انتخاب متغیر گروهی، انتخاب می‌شود. مدل با رگرسیون لجستیک لاسو با خروجی دودویی برازش می‌شود.

نتایج اعمال رگرسیون لاسو لجستیک به‌دست آمده است. ۱۸ تا از بین ۴۰۷ SNPها به‌عنوان مشارکت کننده با بیماری ویروس نیل غربی مشخص شده‌اند. لاسو بقیه ضرایب SNP علامت‌گذار بی‌ربط را صفر قرار می‌دهد. ضرایب مربوط به SNPهای علامت‌گذار در جدول ۱۱.۴ داده شده‌اند.

برآوردگرهای ضرایب همه SNPهای علامت‌گذار بسیار کوچک هستند، در مقیاس ۰٫۱، که حاکی از اثرهای کوچک SNPها بر بیماری ویروسی نیل غربی می‌باشد. rs1044240 که بر ژن ALCAM (مولکول الحاقی سلول گلبول سفید فعال) قرار می‌گیرد دارای ضریب ۰٫۰۹۱ است، با گرفتن $\exp(۰٫۰۹۱) = ۱٫۰۹۵$، داشتن یک آلل جزئی a درنتیجه افزایش در احتمالات ابتلای بیماری ویروسی نیل غربی به $\exp(۰٫۰۹۱) = ۱٫۰۹۵$ بار بیشتر از عدم ابتلا به آن است.

[12] Case

جدول ۱۱.۴: برآوردهای پیشگو توسط لجستیک لاسو

SNP	سمبل ژن	ID مکان	کروموزوم	موقعیت کروموزوم	برآورد
rs10036567	KIAA0141	۹۸۱۲	۵	۱۴۱۳۰۷۸۳۳	−۰/۱۳۹۴۹۷۱
rs1007863	PARVB	۲۹۷۸۰	۲۲	۴۴۳۹۵۴۵۱	۰/۰۱۳۴۸۵۷۲
rs10129889	DYNC1H1	۱۷۷۸	۱۴	۱۰۲۵۰۸۰۵۶	−۰/۱۹۵۱۷۱۷
rs10131813	HOMES	۵۷۵۹۴	۱۴	۲۳۷۴۵۵۳۳	−۰/۰۵۶۹۴۷۰۸
rs10282929	EEF1D	۱۹۳۶	۸	۱۴۴۶۸۱۷۷۷	۰/۰۰۸۴۶۹۰۵۹
rs1031257	LOC100420743	۹۸۴۲	۶	۱۲۷۹۴۵۵۹۴	۰/۰۸۵۷۱۱۴۶
rs10403787	C3PI	۳۸۸۵۰۳	۱۹	۱۰۱۶۶۳۷۵	۰/۰۹۸۰۸۵۷۶
rs10407445	ZIM3	۱۱۴۰۲۶	۱۹	۵۷۶۴۹۹۶۲	−۰/۰۹۸۵۱۴۹
rs10410943	ACTL9	۲۸۴۳۸۲	۱۹	۸۸۰۸۹۰۰	۰/۰۲۲۴۶۶۳۳
rs1043620	HSPA1L	۳۳۰۵	۶	۳۱۷۸۳۷۵۵	−۰/۲۴۲۲۲۷۶
rs1044240	ALCAM	۲۱۴	۳	۱۰۵۲۵۸۸۶۱	۰/۰۹۱۳۰۶۸۶
rs10445686	RAB3GAP1	۲۲۹۳۰	۲	۱۳۵۸۹۳۳۷۲	−۰/۰۱۶۰۷۴۵۴
rs1046480	TMEM159	۵۷۱۴۶	۱۶	۲۱۱۸۵۳۸۴	۰/۰۷۲۲۲۱۴
rs10466026	CDH23	۶۴۰۷۲	۱۰	۷۳۵۵۰۹۶۹	۰/۰۴۱۴۹۶۵
rs1046844	SLC25A26	۱۱۵۲۸۶	۳	۶۶۴۲۸۲۸۲	۰/۰۵۹۵۳۳۰۲
rs1048369	GPC4	۲۲۳۹	X	۱۳۲۴۳۷۳۳۷	−۰/۰۲۲۳۷۲۳۷
rs1050348	LAMA4	۳۹۱۰	۶	۱۱۲۴۹۳۸۷۲	−۰/۰۵۵۳۸۱۴۳

rs1043620 که در ژن HSPA1L (شوک حرارتی 70kDa پروتئین ۱ـشبیه) قرار می‌گیرد دارای ضریب ۰/۲۴− است، با گرفتن $\exp(-۰/۲۴)$، داشتن یک آلل جزئی "a" در نتیجه کاهش در احتمالات ابتلا بیماری ویروسی نیل غربی به $\exp(-۰/۲۴) = ۰/۷۹$ با بیشتر از عدم ابتلا آن می‌باشد.

۶.۴ نتیجه‌گیری و آینده تحقیق

در این پایان‌نامه، تعدادی از روش‌ها برای برخورد با مسائل یادگیری آماری که در بعد بسیار بالا قرار می‌گیرند، مطرح شد. این روش‌ها و نکات برجسته شان به‌صورت زیر فهرست می‌شوند:

بوت‌استرپ و اعتبارسنجی متقابل روش‌های بسیار پرقدرتی هستند، به‌ویژه وقتی داده‌های زیادی دردسترس هستند.

انتخاب مدل کلاسیک مطمئنا دقیق‌ترین روش است اما متاسفانه برای بیشتر از ۳۰ متغیر بسیار کند است.

رگرسیون ریج همه متغیرها را تا حد معینی منقبض می‌کند اما هیچ متغیری صفر نمی‌شود. این مدل برای مدل‌هایی که محقق اعتقاد دارد هر متغیر بر پاسخ کمی اثر دارد، مناسب است.

لاسو همزمان برازش و انتخاب مدل را با منقبض کردن ضرایب بعضی از متغیرها به‌سمت صفر، انجام می‌دهد. متاسفانه، به ضرایب بزرگتر اریبی نیز وارد می‌کند. این مدل برای مدل‌هایی که محقق باور دارد بیشتر متغیرها اثری بر پاسخ ندارند، مناسب است.

SCAD شبیه لاسو کار می‌کند اما مسئله اریبی لاسو را حل کرده است. با این حال، یک مسئله برنامه‌ریزی خطی غیر محدب است و از این رو، مستلزم مقداری سختی برای اجرای آن است.

رگرسیون الاستیک نت در موقعیت‌هایی مفید است که محقق باور دارد بسیاری از متغیرها بر پاسخ اثر نمی‌گذارند و آن‌هایی که اثرگذار هستند، عضوی از گروه متغیرهای همبسته هستند.

لاسو گروهی و بریج گروهی کاستی‌هایی دارند: لاسو گروهی قادر به انتخاب متغیر در سطح مجزا نیست و ضرایب بزرگ را به‌شدت منقبض می‌کند. بریج گروهی نیز مشکلات کاربردی را متحمل می‌شود زیرا همه‌جا مشتق‌پذیر نیست. علاوه بر این، هر دو روش، پذیره‌های انعطاف‌ناپذیری را می‌سازند که می‌توانند باعث ایجاد روش‌هایی شوند که وقتی گروه‌ها نامشخص باشند یا به‌صورت تُنُک نشان داده شده باشند، دچار مشکل شوند. نتایج این پایان‌نامه، می‌توان به سایر جریمه‌هایی که در ادبیات تحقیق مطرح شده است، تعمیم داد. می‌توان نسخه گروهی سایر جریمه‌ها را تعریف کرد و با یافتن الگوریتم‌ها، پاسخی برای آن مسئله بهینه‌سازی یافت.

مراجع

[۱] آرست، م. (۱۳۹۵). مقایسه رفتار برخی برآوردگرهای انقباضی بریج در مدل رگرسیونی چندگانه، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود.

[۲] کاظمی، م. (۱۳۹۷). انتخاب متغیر و تشخیص ساختار در مدل‌های نیمه پارامتری، رساله دکتری، دانشگاه صنعتی شاهرود.

[۳] گودرزی، و. (۱۳۹۷). نسخه‌ای استوار از برآوردگر ریج، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود.

[۴] نوروزی‌راد، م. (۱۳۹۶). برآوردگرهای بهبودیافته در برخی مدل‌های رگرسیونی خطی جریمه‌شده، رساله دکتری، دانشگاه صنعتی شاهرود.

[5] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**(6), 716-723.

[6] Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction, *Technometrics*, **16**(1), 125-127.

[7] Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations, *Journal of the American Statistical Association*, **96**, 939-967.

[8] Bakin, S. (1999). Adaptive regression and model selection in data mining problems, Master thesis, The Australian National University.

[9] Bickel, P. J. and Levenia, E. (2008). Regularized estimation of large covariance matrices, *The Annals of Statistics*, **36**(1), 199-227.

[10] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*, Cambridge University Press.

[11] Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection, *Statistics and lts lnterface*, **2**(3), 369.

[12] Breiman, L. (1995). Better subset regression using the nonnegative garrote, *Technometrics*, **37**(4), 373–384.

[13] Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**(6), 109-135.

[14] Breiman, L. and Spector, P. (1992). Sub-model selection and evaluation in regression: the X-random case, *International Statistical Review*, **60**(3), 291-319.

[15] Buhlman, P. and Van de Geer, S. (2011). *Statistics for High-dimensional Data: Method, Theory and Application*, Springer.

[16] Cai, T. T. (2001). Discussion of regularization of wavelet approximations (by A. Antoniadis and J. Fan), *Journal of the American Statistical Association*, **96**, 960–962.

[17] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics*, **35**(6), 2313- 2351.

[18] Cavanaugh, J. E. (2009). Lecture notes on the Bayesian information criterion, Department of Statistics and Actuarial Science, The University of Iowa.

[19] Donoho, D. L. and Johnstone, I.M. (1994). Ideal spatial adaptation by Wavelet shrinkage, *Biometrika*, **81**, 425-455.

[20] Efron, B. (1979). Bootstrap methods: Another look at the Jackknife, *The Annals of Statistics*, **7**(1), 1-26.

[21] Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation, *Journal of the American Statistical Association*, **78**(382), 316-331.

[22] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., (2004). Least angle regression, *The Annals of Statistics*, **32**(2), 407–499.

[23] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American statistical Association*, **96**(456), 1348–1360.

[24] Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery, *Proceedings of the International Congress of Mathematicians, Madrid, Spain*, 595-622

[25] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

[26] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional variable selection: beyond the linear model, *Journal of Machine Learning Research*, **10**, 2013-2038.

[27] Fan, J., and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics*, **23**, 928-961.

[28] Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**(2), 109–135.

[29] Fu, W.J. (1998). Penalized regressions: The bridge versus the LASSO, *Journal of Computational and Graphical Statistics*, **7**(3), 397-416.

[30] Furnival, G. M. (1971). All possible regressions with less computation, *Technometrics*, **13**(2), 403-408.

[31] Furnival M. and Wilson, R.W. (1974). Regressions by leaps and bounds, *Technometrics*, **16**(4), 499-511.

[32] George, E. and Foster, D. (2000). Calibration and empirical bayes variable selection, *Biometrika*, **87**, 731-747.

[33] George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association*, **88**, 881-889.

[34] Golub, G.H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, **21**(2), 215-223.

[35] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall.

[36] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd Edition Springer.

[37] Hocking, R. R. (1976). The analysis and selection of variables in linear regression, *Biometrics*, **32**(1), 1-49.

[38] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**(1), 55–67.

[39] Huang, J., Ma, S., Xie, H. and Zhang, C. H. (2009). A group bridge approach for variable selection, *Biometrika*, **96**(2), 339–355.

[40] Kass, R. E. and Raftery, A. E. (1995). Bayes Factors, *Journal of the American Statistical Association*, **90**(430), 773-795.

[41] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy prediction and model estimation, *International Joint Conference on Artificial Intelligence (IJCAI)*, 1137-1143.

[42] Mallows, C. L. (1973). Some comments on CP, *Technometrics*, **15**(4), 661-675.

[43] Meier, L., Van de Geer, S., Buhlmann, P. (2009). High-dimensional additive modeling, *The Annals of Statistics*, **37**(6B), 3779–3821.

[44] Meinshausen, N. and Buehlmann, P. (2010). Stability selection, *Journal of the Royal Statistical Society, Series B* , **72**(4), 417-473.

[45] Miller, A. (2002). *Subset Selection in Regression*, Chapman and Hall/CRC.

[46] LaMotte, L. R. and Hocking, R. R. (1970). Computational efficiency in the selection of regression variables, *Technometrics*, **12**(1), 83-93.

[47] Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression, *The Annals of Statistics*, **34**(5), 2272-2297.

[48] Lokhorst, J. (1999). The lasso and generalised linear models, Honors Project, The University of Adelaide, Australia.

[49] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(5), 1009-1030.

[50] Roth, V. (2004). The generalized LASSO, *IEEE transactions on neural networks*, **15**(1), 16-28.

[51] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461-464.

[52] Shen, X. and Ye, J. (2002). Adaptive model selection, *Journal of the American Statistical Association*, **97**(457), 210-221.

[53] Shevade, S. K. and Keerthi, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression,*Bioinformatics*, **19**(17), 2246-2253.

[54] Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**(2), 111-147.

[55] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-vbalidation and Akaike's criterion, *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 44-47.

[56] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

[57] West M., Blanchette, C., Dressman, H., Huang, E., Ishida, S. Spang, R., Zuzan, H., Olson Jr, J. A., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer using gene expression, *Proceedings of the Natioanl Academy of Sciences of the United States of America*, **98**, 11462-11467.

[58] Wu, Y., Boos, D. D. and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables, *Journal of the American Statistical Association*, **102**(477), 235-243.

[59] Xie, J. and Zeng, L. (2010). Group variable selection methods and their applications in analysis of genomic data, In Frontiers in Computational and Systems Biology, 231–248.

[60] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49-67.

[61] Zhang, T. (2008). Adaptive forward-backward greedy algorithm for sparse learning with linear models, Neural Information Processing System.

[62] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of statistics*, **38(2)**, 894–942.

[63] Zhang, C. H. and Huang, J . (2008). Model selection consistency of the lasso in high dimensional linear regression, *The Annals of Statistics*, **36**, 1567-1594.

[64] Zhao, P. and Yu, B. (2006). On model selection consistency of the LASSO, *The Journal of Machine Learning Research*, **7**, 2541-2563.

[65] Zhao, P., Rocha, G. and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties, *The Annals of Statistics*, **37**(6), 3468-3497.

[66] Zou, H. (2006). The adaptive LASSO and its oracle properties, *Journal of the American Statistical Association*, **101**(476), 1418-1429.

[67] 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.

[68] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the LASSO, *The Annals of Statistics*, **35**(5), 2173-2192.

[69] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models, *Annals of Statistics*, **36**(5), 1509-1533.

Abstract

Nowadays, due to the development of technology, it has been possible to store and analyze data with a large number of variables: but it should be noted that if the recorded variables are not significant or the number of them is high, classical estimation techniques are ineffective and in addition to identifying and removing redundant variables, non-classical methods for estimating model parameters should be used. In the regression modeling, by using penalized estimators, we are able to select both significant variables and estimate parameters. In this regard, penalized group estimators can be used if the variables are grouped in such a way that all of them have to be eliminated or they should all remain in the model. In this dissertation, we briefly review some of the most common and applicable modern variable selection methods and investigate penalized group estimators via a series of numerical studies, simulations, and analysis of genomic real data and their use in logistic regression models.

Key words: Genomics data, Group LASSO, High dimensional, SCAD, Variable selection.

Shahrood University of Technology

Faculty of Mathematical Sciences

MSc Thesis in: Mathematical Statistics

On Variable Selection Methods in High Dimensional Regression Models

By: Seyedeh Faeze Mirsalari

Supervisor

Dr. Mohammad Arashi

Advisor

Dr. Mina Norouzirad

September 2019