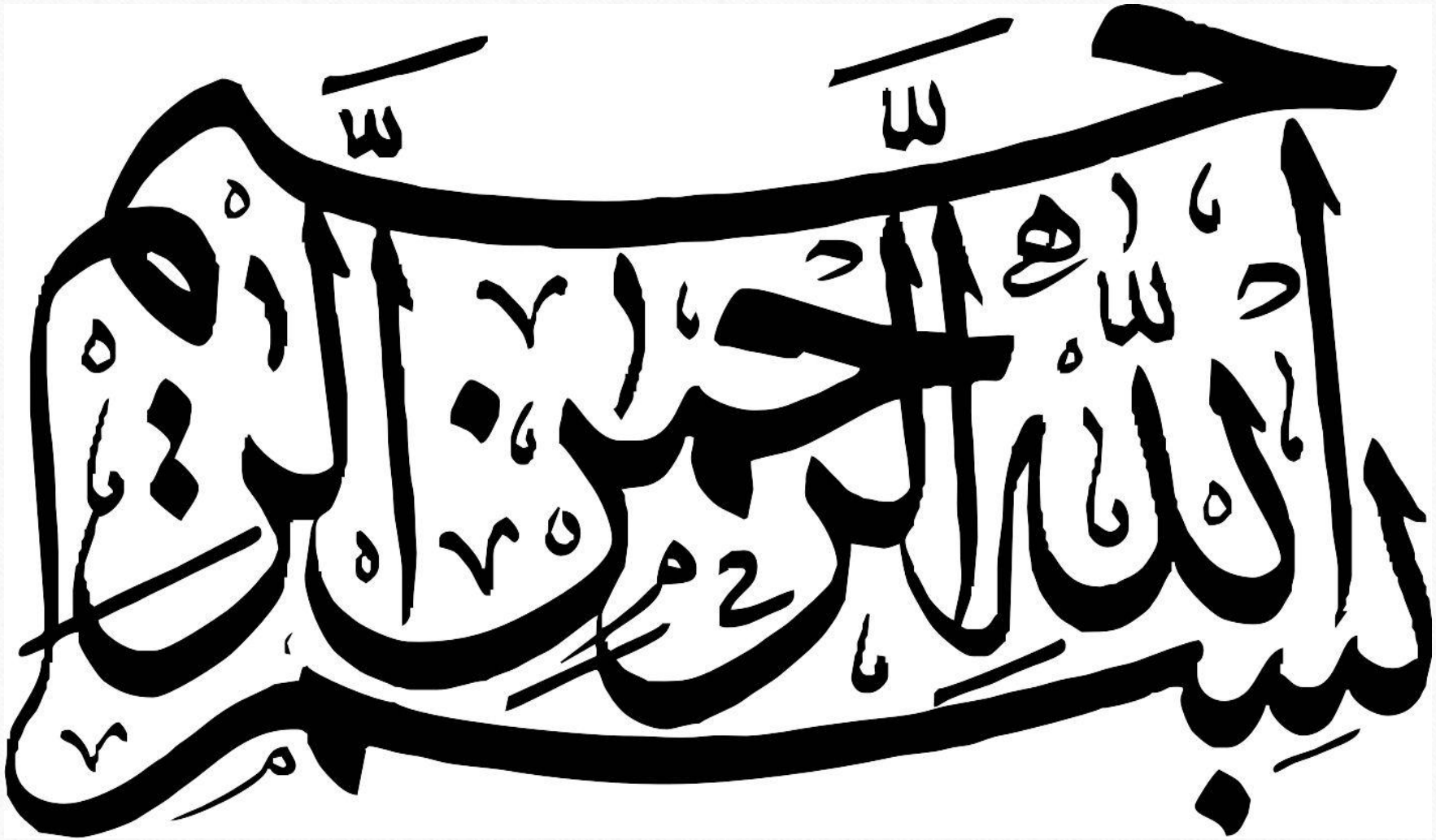

بسم الله الرحمن الرحيم

نسرین ندافی
طوبی رضائی

درس مدل های خطی (۱)- shrinkage methods

مقدمه:

تحلیل رگرسیونی روشی آماری برای بررسی و به مدل در آوردن ارتباط بین متغیرها است. در بسیاری از مسائل از جمله پزشکی تعداد متغیرها بسیار زیاد و تعداد مشاهدات کم هستند که به مسئله n کوچک و p بزرگ معروف است. این مسئله مشکلات بسیار زیادی در برازش مدل های رگرسیون به وجود می آورد.

از طرفی بسیاری از متغیرها رفتار یکسانی با متغیرهای دیگر دارند یا در واقع برخی از متغیرها ترکیب خطی از یک یا چند متغیر دیگر هستند. از این رو، در رویارویی با این گونه مسائل، زیر مجموعه کوچکی از متغیرها که دارای بیش ترین تاثیر بوده راانتخاب و به برآورد آن ها پرداخته می شود. بنابراین انتخاب متغیر و برآورد ضرایب، اساسی ترین بخش در مدل سازی رگرسیونی است.

روشهای برآوردیابی کمترین توان های دوم، انتخاب متغیر به صورت پیشرو و غیره، در مواجه با داده هایی که از ویژگی های متفاوتی برخوردار باشند، عملکرد قابل اطمینانی از خود نشان نمی دهند. از آسیب های مدل در هنگام استفاده از این روش ها می توان به عدم پایداری، دقت پیش بینی کم و انتخاب نادرست متغیرها، اشاره نمود. به علاوه این مشکلات زمانی که همبستگی بین متغیرهای پیشبین زیاد باشد، تشدید نیز می شوند. روش های انقباضی به عنوان راهکاری برای کاهش این مشکلات به خصوص زمانی که همبستگی بین متغیرهای پیشبین زیاد باشد، مورد توجه قرار گرفته اند.

روش های انتخاب متغیر و تخمین پارامترهای مدل

اهمیت انتخاب متغیر:
چرا ما نمی‌خواهیم همه متغیرهای توضیحی در مدل باشد؟

- مدل واقعی معمولا بسیار پیچیده است.
- مدل کامل و درست مجهول و ناشناخته است.
- وجود متغیرهای اضافی در مدل باعث بیش برازش می گردد.

مدل های با تعداد متغیر کم همواره دلخواه تحلیل گران آماری بوده است. زیرا:
۱- روابط بین متغیرهای علمی را به صورت ساده بیان کرده و قابلیت تفسیر ساده تری دارد.
۲- هزینه های محاسباتی و صرف زمان زیاد روش های انتخاب متغیر را با چالش بسیار جدی در زمینه استفاده از داده ها با بعد بالا رو برو کرده است.
۳- انتخاب یک مدل آماری خوب مدلی است که استفاده از آن آسان و قابل فهم و قابل توضیح برای دیگران است.

معیارهای انتخاب متغیر:

۱.آزمون فرض:

معیارها زمانی قابل استفاده هستند که همه زیر مدل های ممکن شناخته شده باشند.

۲.روشهای انقباضی

برخی از معیارها فقط زمانی کاربرد دارند که مدل مورد نظر بر حسب پارامترها خطی باشد.

۳.روشهای بیزی

محاسبه این معیارها، برای تمام زیر مدل های ممکن، محاسبات زیادی دارد. بخصوص اگر متغیرهای تصادفی زیاد باشد.

استفاده از آزمون فرض:

انتخاب پیشرو

حذف پسرو

روش گام به گام

در مدل رگرسیون خطی $y = x\beta + \varepsilon$ هدف براورد پارامتر بتا بود که براساس داده های مشاهده شده براورد می شوند.

براورد بتا به روش حداقل مربعات معمولی (ols)

$$\hat{\beta} = \left(x^T x\right)^{-1} x^T y$$

در این حالت مشکلاتی بوجود می آید. از جمله :

- وقتی $n < p$ باشد، پارامترها نمی توانند با استفاده از براورد بالا به صورت یکتا تعریف شوند.
- در مواردی که متغیرها وابستگی دارند، براورد بالا باعث بزرگی واریانس می شوند. برآوردگر کارایی لازم ندارد.

مسئله هم خطی و کاربردی نبودن برآوردگر کمترین توان های دوم، منجر به توسعه برآوردگرهای اریب مانند برآوردگرهای انقباضی استاین، مولفه های اصلی، کمترین توان های دوم جزئی، نوع لیو، تقریبا نااریب، نوع ریج و لاسو که برآوردگرهای بهبودیافته را ارائه می دهند، شده است. در دو دهه اخیر روش های انقباضی مختلفی برای برآورد ضرایب رگرسیونی ارائه شده است، که در بین آن ها روش انقباضی بریج از توجه شایانی برخوردار شده است.

استفاده از روش انقباضی

روش انقباضی:
هنگامی این روش استفاده می شود که تعداد متغیرهای مستقل زیاد و وابستگی بین متغیرهای نیز زیاد باشد.
نحوه ی براورد: در نظر گرفتن مقداری اریبی برای براوردگرها در تلاش برای کاهش واریانس است به طوری که در نهایت میانگین مربع خطا کاهش یابد.
در این روش عمل براوریابی و انتخاب متغیر هم زمان صورت می گیرد.

- رگرسیون مرزی یا ریج
- لاسو
- لارس
- الاستیک نت

روش انقباضی بریج:

فرانک و فریدمن مجموع توان دوم باقیمانده RSS را تحت قید $\sum_{j=1}^{p}\left|\beta_j\right|^{\gamma} \le t$ به ازای $\gamma \ge 0$ کمینه کردند و آن را به عنوان تعمیمی از رگرسیون ریج به نام رگرسیون بریج برای مقابله با هم خطی ارائه نمودند.

براوردگر حاصل به صورت زیر می باشد:

که در آن $\lambda, t \ge 0$ پارامترهای کنترل بوده و میزان انقباض تحمیل شده به ضرایب را کنترل می کند.

$$\hat{\beta}_{Bridge} = \arg\min\left\{(Y - X\beta)^T (Y - X\beta) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|^{\gamma}\right\}$$

رگرسیون مرزی(ریج)

هارلوت و کنارد در سال ۱۹۷۰ پیشنهاد کردند که ناپایداری براوردگر $\hat{\beta}_{OLS} = \left(x^T x\right)^{-1} x^T y$ به وسیله اضافه کردن یک مقدار ثابت و کوچک لاندا به عناصر قطری ماتریس $x^T x$ قبل از معکوس کردن بهبود می یابد که نتیجه آن براوردگر رگرسیون مرزی شد.

$$\hat{\beta}_{Ridge} = \arg\min\left\{\|y - x\beta\|^2 + \lambda\|\beta\|^2\right\}$$

رگرسیون مرزی:

براوردگر رگرسیون مرزی را می توان به فرم زیر نوشت:

$$\hat{\beta}_{Ridge} = \arg\min\left\{\|y - x\beta\|^2 + \lambda\|\beta\|^2\right\} = \left(x^T x + \lambda I_P\right)^{-1} X^T Y$$

که در آن لاندا پارامتر کنترل است.

لاندا معمولا به گونه ای انتخاب می‌شود که براوردگر رگرسیون مرزی میانگین مربعات خطای کمتری نسبت به براوردگر حداقل مربعات داشته باشد.

این براوردگر در مقایسه با براوردگر کمترین مربعات خطا $\hat{\beta}_{OLS} = \left(x^T x\right)^{-1} x^T y$ اریب است اما واریانس کمتری دارد.

این براوردگر معمولا در مواردی با تعداد دامنه های بالا (تعداد متغیرهای مستقل زیاد) کاربرد دارد.

رگرسیون مرزی یک فرم خاص از محدودیت را روی پارامترهای بتا قرار می دهد. β_{Ridge} به گونه ای به دست آمده است که مجموع مربعات زیر را مینیمم کند.

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j{}^2$$

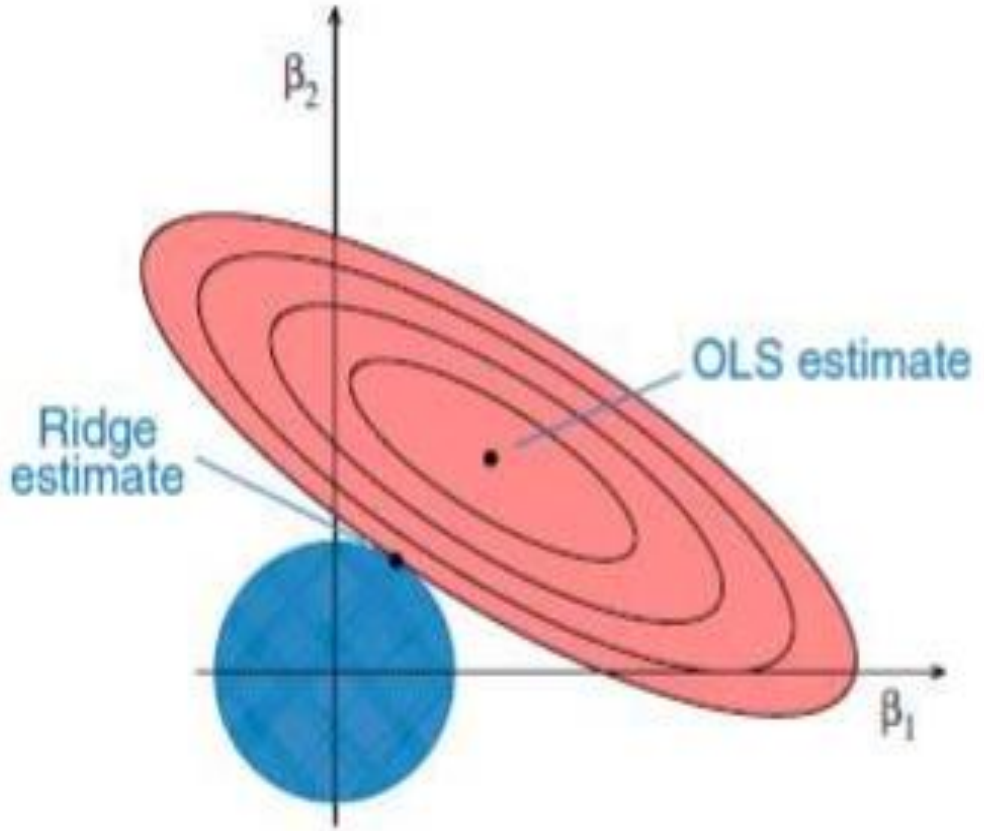
که معادل min کردن $\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2$ به شرط آن است که $\sum_{j=1}^{p}\beta_j{}^2<c$ برای $c>0$

بنابراین رگرسیون مرزی ، محدودیت هایی را روی پارامترهای بتا در مدل خطی به وجود می آورد. در این مورد کاری که ما انجام می دهیم این است که به جای min کردن مجموع مربعات خطا ما یک عبارت تاوان روی بتا ها نیز داریم این عبارت تاوان لاندا برابر مربع نرم بردار بتا است. این به این معنی است که اگربتاها مقادیر بزرگی اختیار کنند تابع بهینه تاوانیده می شود. ما ترجیح می دهیم که بتاها مقادیرکوچک بگیرند یا این که به صفر نزدیک شوندکه عبارت تاوان کوچک شود.

$$\lambda \sum_{j=1}^{p} \beta_j{}^2$$

از دید بهینه سازی، عبارت تاوان معادل محدودیتی روی بتاهاست. تابع همچنان جمع مربعات خطا است اما با این تفاوت که ما محدودیت این که نرم بتاها کمتر از C باشد را اعمال کرده ایم.
ارتباطی بین لاندا و C وجود دارد. لاندا بزرگتر باعث می شود که ترجیح دهیم بتاها نزدیک صفر باشد.
در مورد خاص وقت لاندا صفرباشد ما به سادگی می توانیم رگرسیون خطی ساده را انجام دهیم و در موردی که لاندا به سمت بی نهایت میل می کند ما همه بتاها را صفر قرار می دهیم.

Geometric Interpretation of Ridge Regression:



بیضی با RSS مطابقت دارند و داخلی ترین بیضی کمترین RSS را دارد و RSS، در برآورد ols مینیمم می شود.
برای P=2، محدودیت در رگرسیون مرزی با دایره مطابقت دارد.

ما سعی می کنیم که سایز بیضی ها و دایره ها را به طور همزمان کوچک کنیم . برآورد مرزی نقطه ای است که بیضی ها و دایره ها به هم مماس اند.

یک مبادله بین عبارت تاوان و RSS وجود دارد . ممکن است که یک بتا بزرگ، جمع مربعات معتبری داشته باشد اما باعث افزایش عبارت تاوان می شود. این به همین دلیل است که ما بتاهای کوچکتر را در مقابل جمع مربعات خطای بدتر ترجیح می دهیم.

ایرادات وارده به روش رگرسیون مرزی:

این براوردگرها اریب هستند.

مانع از صفر برآورد شدن ضرایب رگرسیونی می شود.

اگر $X^T X = nI_P$ بنابراین $\hat{\beta}_{Ridge} = \frac{n}{n+\lambda}\hat{\beta}_{OLS}$

در این مورد برآوردگر مرزی همیشه انقباض به سمت صفر را تولید می کند و لاندا مقدار انقباض را کنترل می کند.

$$\hat{\beta}_{Ridge} = \arg\min\left\{\|y - x\beta\|^2 + \lambda\|\beta\|^2\right\} = \left(x^T x + \lambda I_P\right)^{-1} X^T Y$$

Lasso:

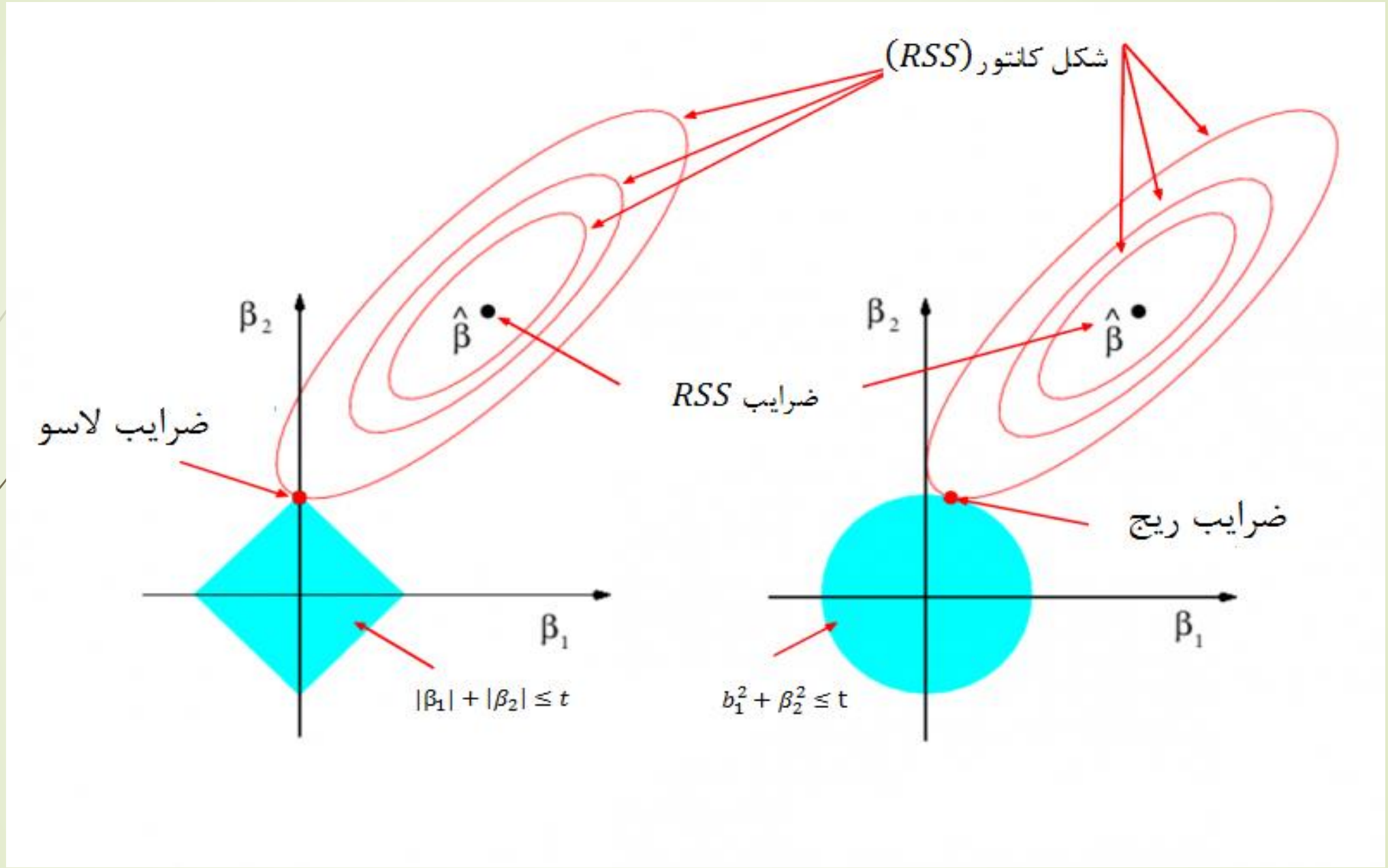
این روش شبیه به روش رگرسیون مرزی است با این تفاوت که به جای استفاده از تابع تاوان درجه دو از تابع تاوان قدر مطلق ضرایب استفاده می کند و باعث می شود که بعضی از ضرایب دقیقا به صفر کاهش یابند.

$$\beta_{Lasso} = \arg\min \left\{ \|y - x\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

مزایای روش لاسو:

روش لاسو به لحاظ پایداری و دقت پیش بینی برآوردگرها عملکرد قابل قبولی از خود نشان داده است.

نسبت به رگرسیون مرزی اکثر مواقع دقت پیش بینی بالایی دارد.


شکل کانتور (RSS)
β_2
$\hat{\beta}$
ضرایب لاسو
β_1
$|\beta_1| + |\beta_2| \leq t$
β_2
$\hat{\beta}$
ضرایب RSS
ضرایب ریج
β_1
$b_1^2 + \beta_2^2 \leq t$

مقایسه روش رگرسیون لاسو و روش رگرسیون مرزی :

در شکل وقتی که p افزایش یابد تعداد ضلع های شکل افزایش می یابد و شانس صفر شدن ضرایب زیادتر می شود. بنابراین SS ها، انقباض و انتخاب زیر مجموعه را به طور همزمان انجام می دهند.

رگرسیون مرزی برآوردگر اریب تولید می کند اما واریانس را کاهش می دهد.

رگرسیون مرزی همه ضرایب را به یک مقدار غیر صفر کاهش می دهد اما در روش لاسو مقداری از ضرایب دقیقا به صفر کاهش می یابد.

معایب روش لاسو:

محاسبات بسیار پیچیده

فاقد ویژگی نااریبی است.

یک مقدار تاوان ثابت برای ضرایب در نظر میگیرد.

عدم ثبات کافی در انتخاب متغیرهای موثر زمانی که داده ها شامل گروه هایی از متغیرهای پیش بینی به شدت به هم وابسته هستند.

روش الاستیک نت:

برآورد ضرایب در این روش به فرم زیر است:

$$\hat{\beta}_{NEN} = \arg\min\left\{\|y - x\beta\|^2\right\} + \lambda\sum_{j=1}^{p}|\beta_j| + \lambda(1-\alpha)\sum_{j=1}^{p}\beta_j{}^2$$

این فرم از تابع تاوان علاوه بر توانایی صفر براورد کردن برخی از ضرایب، توانایی برابر براورد کردن ضرایب متغیرهایی که اثر یکسانی روی متغیر پاسخ دارند یا بشدت همبسته هستند نیز دارد.

مزایای روش الاستیک نت:

۱. این روش هم از نظر دقت پیش بینی و هم از نظر انتخاب مدل صحیح از دیگر روش های مذکور بهتر عمل می کند.
۲. این روش از نظر تمایز بین متغیرهای مؤثر وغیرمؤثر همبسته نیز از دیگر روش های مذکور بهتر عمل می کند.

حال کارایی برآوردگرهای پیشنهاد شده در یک مثال واقعی مورد مطالعه قرار می گیرد. برای این منظور از داده های پروستات که استیمی و همکاران (۱۹۸۹) ارائه داده اند، استفاده شده است. این داده ها وابستگی بین آنتیژن نوع خاص پروستات و اندازه کلینیکی مردانی که به پروستاتکومی مراجعه کرده اند را نشان می دهند.

این نتایج نشان می دهد زمانی که تعداد متغیرها کمتر از مشاهدات است یا وقتی داده ها دارای هم خطی می باشند، برآورگر کمترین توان های دوم نمی تواند به خوبی ضرایب را به دست آورد و مقدار میانگین توان دوم خطای آن زیاد است. درحالی که، برآوردگر لاسو بهتر از برآوردگر کمترین توان های دوم ضرایب را برآورد کرده و دارای میانگین توان دوم خطای کمتری است. در مجموع، رگرسیون ریج دارای بهترین دقت پیش بینی در بین سایر روش هاست و عیب اصلی این روش غیر صفر برآورد کردن ضرایب و عدم حذف متغیر از مدل است.

نتیجه گیری:

۱ . باید توجه داشت که مشخص کردن ساختار ماتریس کواریانس متغیرهای مستقل برای انتخاب درست یک روش برآوردیابی و انتخاب متغیر می تواند مفید باشد.

۲ . برای مدل هایی با تعداد متغیرهای متوسط یا کم روش الاستیک نت هم در حالتی که همبستگی بین متغیر ها زیاد باشد و هم در حالتی که همبستگی چندان قابل توجه نباشد عملکرد قابل قبولی از خود نشان داده است.

۳ . البته هم در حالتی که همبستگی بین متغیر ها زیاد نباشد می توان از روش های دیگری مانند لاسو و لارس استفاده کرد. از نظر محاسبات ساده نیز می باشد.

منابع:

-An Introduction to statistical learning G.James, D.witten, T.Hastie,R.Tibshiren.

- Data mining and analysis Jonathan Taylor.

- Shrinkage regression Rolf sundbery.


با تشکر از توجه شما