بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

پایان‌نامه کارشناسی ارشد آمار ریاضی

رگرسیون جریمه شده در بعد بالا با تابع جریمه لاسوی مشتق پذیر

نگارنده: سمانه رجبلو

استاد راهنما

دکتر محمد آرشی

شهریور۱۳۹۷

شماره:
تاریخ:

بأسمه‌تعالی

دانشگاه صنعتی شاهرود

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 **سمانه رجبلو** با شماره دانشجویی ۹۵۰۶۵۰۴ رشته **آمار** گرایش **آمار ریاضی** تحت عنوان: **رگرسیون جریمه شده در بعد بالا با تابع جریمه لاسوی مشنق پذیر** که در تاریخ **۱۳۹۷/۰۶/۱۳** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

قبول (با درجه: عالی) ☑ **مردود** ☐

نوع تحقیق: **نظری** ☑ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبه علمی	امضاء
۱ـ استادراهنمای اول	دکتر محمدآرشی	دانشیار	
۲- استادراهنمای دوم	------	------	------
۳- استاد مشاور	------	------	------
۴- نماینده تحصیلات تکمیلی	دکتر داود شاهسونی	دانشیار	
۵- استاد ممتحن اول	دکتر حسین باغیشنی	استادیار	
۶ـ استاد ممتحن دوم	دکتر محمدرضا ربیعی	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 به

آنان که مهر آسمانی‌شان آرام بخش آلام زمینی‌ام است به استوارترین تکیه‌گاهم، دستان پر مهر پدرم به زیباترین نگاه زندگیم، چشمان زیبای مادرم که هر آنچه آموختم در مکتب عشق شما آموختم و هر چه بکوشم قطره‌ای از دریای بی کران مهربانیتان را سپاس نتوانم بگویم.
امروز هستی‌ام به امید شماست و فردا کلید باغ بهشتم رضای شما باشد که حاصل تلاشم نسیم گونه غبار خستگیتان را بزداید.

بوسه بر دستان پر مهرتان

سپاس‌گزاری

به مصداق (من لم یشکر المخلوق لم یشکر الخالق) بسی شایسته است از استاد فرهیخته و فرزانه جناب آقای دکتر محمد آرشی که با کرامتی چون خورشید، سرزمین دل را روشنی بخشیدند و گلشن سرای علم و دانش را با راهنمایی‌های کارساز و سازنده بارور ساختند تقدیر و تشکر نمایم.
از سرکار خانم دکتر مینا نوروزی‌راد به دلیل یاری‌ها و راهنمایی‌های بی چشم‌داشت ایشان که بسیاری از سختی‌ها را برایم آسان‌تر نمودند، سپاسگزارم.
همچنین از پدر و مادر عزیز، دلسوز و مهربانم که آرامش روحی و آسایش فکری فراهم نمودند تا با حمایت‌های همه جانبه در محیطی مطلوب، مراتب تحصیلی و نیز پایان‌نامه درسی را به نحو احسن به اتمام برسانم، سپاسگزاری نمایم.

پروردگارا حسن عاقبت، سلامت و سعادت را برای آنان مقدر نما.

سمانه رجبلو
شهریور۱۳۹۷

تعهد نامه

اینجانب **سمانه رجبلو** دانشجوی کارشناسی ارشد رشته **آمار** دانشکده **علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **رگرسیون جریمه شده در بعد بالا با تابع جریمه لاسوی مشتق پذیر** ، تحت راهنمایی **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.

سمانه رجبلو
شهریور۱۳۹۷

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

در مسائل تحلیل رگرسیون اگر تعداد متغیرهای توضیحی زیاد باشد، برآورد پارامترها به روش کمترین توان‌های دوم راه مناسبی نبوده و باید از برآوردگرهای جریمه‌شده استفاده کرد. یکی از موثرترین و کاراترین برآوردگرهای جریمه‌شده، لاسو است که در صفر مشتق‌پذیر نمی‌باشد و بنابراین توزیع مجانبی آن نرمال نیست. در این پایان‌نامه برآوردگر لاسوی مشتق‌پذیر را مورد بررسی قرار می‌دهیم که دارای توزیع مجانبی نرمال است و خواص خوبی در مقایسه با خیلی از برآوردگرهای جریمه‌شده دارد. با انجام چند مطالعه شبیه‌سازی و تحلیل داده‌های واقعی، عملکرد این برآوردگر و برآوردگر لاسوی مشتق‌پذیر تفاضلی را مورد ارزیابی قرار می‌دهیم.

کلمات کلیدی: برآوردگر لاسوی مشتق‌پذیر، برآوردگر تفاضلی، تابع جریمه، رگرسیون جریمه‌شده در بعد بالا

فهرست مقالات مستخرج از پایان‌نامه

۱. رجبلو، س، آرشی، م. و نوروزی‌راد، م.(۱۳۹۷) کاربرد لاسوی مشتق‌پذیر در تحلیل داده‌های سرطان پرستات، چهاردهمین کنفرانس آمار ایران، شاهرود

۲. Rajabloo, S., Arashi, M. and Norouzirad, M. (2018) Differenced-based differentiable linear models, in progress.

پیش گفتار

در تحلیل مسائل رگرسیونی چنانچه تعداد متغیرهای توضیحی (p) زیاد باشد، تفسیر مدل نهایی ساده نبوده و اغلب در تحلیل مسائل بعد بالای این چنینی سعی می‌شود با استفاده از روش‌های انتخاب متغیر، تعداد متغیرها را کم کرده و متغیرهای معنی‌دار را در مدل نگه داشت. برای این منظور به جای استفاده از روش کمترین توان‌های دوم، از صورت جریمه‌شده آن استفاده می‌کنند و برخی از برآوردگرهای جریمه‌شده حاصل این قابلیت را دارند که مقدار p را کاهش دهند. از این میان برآوردگر لاسو که توسط تیبشیرانی (۱۹۹۶) ارائه شد، علاوه بر برآورد پارامترها همزمان تعداد متغیرها را کاهش داده و انتخاب متغیر نیز انجام می‌دهد. اما خاصیت اوراکل را به دلیل مشتق‌پذیر نبودن در صفر ندارد و لذا توزیع مجانبی آن نرمال نیست. حاصلی مشهدی و وینچوتی (۲۰۱۶) برآوردگر جایگزینی معرفی کردند که مشتق‌پذیر است. در این پایان‌نامه پس از مرور انواع مختلف برآوردگرهای جریمه‌شده که انتخاب متغیر انجام می‌دهند، برآوردگر لاسوی مشتق‌پذیر حاصلی‌مشهدی و وینچوتی را معرفی کرده و خواص آن را از دیدگاه‌های نظری و عددی بررسی می‌کنیم. به عنوان تحقیقی نو، برآوردگر لاسوی مشتق‌پذیر را در مدل تفاضلی، برای تحلیل رگرسیون نیمه‌پارامتری، معرفی کرده و رفتار آن را در قالب یک مثال شبیه‌سازی مورد ارزیابی قرارمی‌دهیم. از این مجموعه دو مقاله به‌دست آمده که در فهرست مقالات مستخرج از پایان‌نامه آورده شده‌اند. این مجموعه شامل ۳ فصل و یک ضمیمه است که محتویات آن‌ها به طور مختصر به قرار زیر است. در فصل ۱ به معرفی برآوردگرهای جریمه‌شده می‌پردازیم که در این میان تمرکز بر روی دو برآوردگر معروف ریج و لاسو است. در فصل ۲ برآوردگر لاسوی مشتق‌پذیر معرفی شده، خواص نظری آن را بررسی کرده و کاربرد آن با استفاده از مطالعات عددی مورد ارزیابی و مقایسه قرار گرفته است. مقاله ۱ مستخرج از این پایان‌نامه نتیجه مطالعات این فصل می‌باشد. در فصل ۳، رگرسیونی نیمه‌پارامتری را هدف گرفته و برآوردگر لاسوی مشتق‌پذیر تفاضلی را معرفی می‌کنیم. نشان می‌دهیم برآوردگر ارائه شده دارای توزیع مجانبی نرمال است. نتایج این فصل از نویسنده این مجموعه و جدید می‌باشد. حاصل یافته‌های این فصل به طور خلاصه در مقاله ۲ مستخرج از این پایان‌نامه آمده است. در آخر ملزومات نظری و کاربردی این مجموعه در پیوست آمده است.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

برآوردگر لاسو

۱.۱ مقدمه

مدل‌بندی و تحلیل رگرسیون یکی از روش‌های آماری برای به‌دست آوردن رابطه‌ی بین یک یا چند متغیر توضیحی یا تبیینی با یک یا چند متغیر پاسخ است، که در بیشتر حوزه‌های علوم مورد استفاده قرار می‌گیرد. مدل رگرسیون خطی ساده به صورت زیر تعریف می‌شود:

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{۱.۱}$$

که در آن Y متغیر وابسته (پاسخ)، X متغیر توضیحی (تبیینی)، β_0 عرض از مبدأ، β_1 شیب خط (ضریب رگرسیونی) و ϵ مولفه خطای تصادفی است. مرسوم است فرض کنیم $E(\epsilon) = 0$ و $E(\epsilon^2) = \sigma^2$، $(\sigma^2 > 0)$. در عمل تعداد متغیرهای توضیحی بیشتر از یکی است؛ به عنوان مثال در تحلیل داده‌های سرطان پروستات علاقه‌مندیم رابطه بین سطح آنتی ژن اختصاصی پروستات[۱] (lpsa) (پادتن مخصوص پروستات) با هشت متغیر دیگر از جمله لگاریتم حجم سرطان[۲] (lcanvol)، لگاریتم وزن پروستات[۳]

[۱] Level of prostate specific antigen

[۲] Log cancer volume

[۳] Log prostate weight

(lweight)، لگاریتم مقدار یاخته پروستات خوش‌خیم[۴] (lbph)، لگاریتم کپسول نفوذی[۵] (lcp)، سن[۶]، میزان گلیسون[۷] (gleason)، درصد گلیسون ۴ یا ۵[۸] (pgg45) و تحاجم بافتی[۹] (svi) را بررسی کنیم. در این مثال مدل (۱.۱) قابل استفاده نبوده و برای مدل‌بندی متغیر پاسخ با متغیرهای توضیحی عنوان شده می‌توان از روش رگرسیون چندگانه استفاده کرد.

مدل رگرسیون خطی چندگانه برای p متغیر توضیحی به صورت کلی زیر است

$$\begin{aligned} y_۱ &= \beta_۰ + \beta_۱ x_{۱۱} + \beta_۲ x_{۱۲} + \cdots + \beta_p x_{۱p} + \epsilon_۱ \\ y_۲ &= \beta_۰ + \beta_۱ x_{۲۱} + \beta_۲ x_{۲۲} + \cdots + \beta_p x_{۲p} + \epsilon_۲ \\ &\vdots \\ y_n &= \beta_۰ + \beta_۱ x_{n۱} + \beta_۲ x_{n۲} + \cdots + \beta_p x_{np} + \epsilon_n, \end{aligned} \qquad (۲.۱)$$

که در آن به ازای $(i = ۱, \ldots, n)$، y_i متغیر پاسخ i امین مشاهده، $\beta_۰$ عرض از مبدأ، β_j، به ازای j $(j = ۱, \ldots, p)$ امین پارامتر رگرسیونی، ϵ_i، i امین مولفه خطای تصادفی است که تغییرات تصادفی y_i را که به وسیله متغیرهای توضیحی مربوطه بیان نمی‌شود را نشان می‌دهد. بدیهی است در این حالت x_{ij} متغیر توضیحی مربوط به مشاهده i ام از متغیر توضیحی j است. بدون کاستن از کلیت مسئله فرض می‌کنیم $\beta_۰ = ۰$. (برای مشاهده این که چطور با مرکزی‌کردن مشاهدات می‌توان اثر عرض از مبدأ را حذف کرد پیوست آ.۱.۰ را ببینید.)

مدل (۲.۱) با نماد ماتریسی به صورت زیر نمایش داده می‌شود

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad (۳.۱)$$

که در آن

$$\boldsymbol{y} = \begin{pmatrix} y_۱ \\ y_۲ \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} x_{۱۱} & x_{۱۲} & \ldots & x_{۱p} \\ x_{۲۱} & x_{۲۲} & \ldots & x_{۲p} \\ \vdots & \vdots & & \vdots \\ x_{n۱} & x_{n۲} & \ldots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_۱ \\ \beta_۲ \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_۱ \\ \epsilon_۲ \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

در حالت کلی فرض می‌شود

[۴] Log benign prostatic hyperplasia
[۵] Log capsular penetration
[۶] Age
[۷] Gleason score
[۸] percentage Gleason scores 4 or 5
[۹] Seminal vesicle invasion

۱. $E(\boldsymbol{\epsilon}) = ۰$.

۲. $E(\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}) = \sigma^۲ I$ ، $\sigma^۲ > ۰$.

۳. ماتریس $\boldsymbol{X}$ غیرتصادفی است.

در مدل (۳.۱) هدف اولیه برآورد بردار $\boldsymbol{\beta}$ است. یکی از روش‌های برآورد بردار $\boldsymbol{\beta}$ روش کمترین توان‌های دوم[۱۰] (LS) است. می‌توان مسئله را به صورت زیر فرمول‌بندی کرد

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_۲^۲. \tag{۴.۱}$$

روش کمترین توان‌های دوم رهیافتی است که محک زیر را کمینه می‌کند

$$\begin{aligned} S(\boldsymbol{\beta}) &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_۲^۲ \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^\top \boldsymbol{y} - ۲\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}. \end{aligned} \tag{۵.۱}$$

از (۵.۱) نسبت به $\boldsymbol{\beta}$ مشتق می‌گیریم، داریم

$$\frac{dS(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = -۲\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

مشتق S را برابر بردار صفر قرار می‌دهیم و یک سیستم از معادلات نرمال به صورت زیر به‌دست می‌آوریم:

$$\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = ۰.$$

چنانچه $\boldsymbol{X}^\top \boldsymbol{X}$ معکوس پذیر باشد، برآوردگر LS بردار $\boldsymbol{\beta}$ عبارت است از

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-۱} \boldsymbol{X}^\top \boldsymbol{y}. \tag{۶.۱}$$

برآوردگر $\hat{\boldsymbol{\beta}}$ ناریب است. این برآوردگر، خواص خوبی داشته و طبق قضیه گوس‌ـ مارکف (کاریا وکوراتا[۱۱]، ۲۰۰۸) در بین تمام برآوردگرهای خطی ناریب به صورت $\hat{\boldsymbol{\beta}}^* = \boldsymbol{A}\boldsymbol{y}$ که در آن A یک ماتریس معلوم $p \times n$ است و دارای کمترین واریانس است.

روش LS ساده است، به آسانی قابل فهم است و برای نوع خاصی از داده‌ها بسیار کاربردی است. اگر مدلی با تعداد متغیرهای کم در مقابل تعداد مشاهدات موجود باشد، روش LS نتایج قابل قبولی ارائه می‌کند. با این وجود، اگر p، (تعداد متغیرهای توضیحی) به اندازه کافی بزرگ باشد و به ویژه، وقتی بزرگتر از n (تعداد مشاهدات)

[۱۰] Least squares

[۱۱] Kariya and Kurata

باشد، برآورد مدل با روش LS بسیار پیچیده است. مدل پیچیده، تعداد متغیرهای زیادی را در برمی‌گیرد که براساس اینکه روش LS معمولا ضرایب غیر صفر تولید می‌کند، تفسیرپذیری دشواری دارد. همچنین، در یک مدل معمولی که انتخاب متغیرها در نظر گرفته نشده است، بیش‌برازشی مشاهده می‌شود. بیش‌برازشی، وقتی اتفاق می‌افتد که یک مدل پیچیده در اختیار داشته باشیم. به عبارت دیگر، تعداد متغیرها در مقایسه با تعداد مشاهدات، بیشتر از حد انتظار وجود داشته باشد که منجر به پیش‌بینی ضعیفی می‌شود زیرا برآورد شدیدا تحت تأثیر خطای داده‌ها قرار می‌گیرد که این روزها یک مسئله‌ی رایج در روش‌های یادگیری آماری است (فریدمن[۱۲]، ۲۰۰۱ و همکاران).

یک روش برای مواجهه با بیش‌برازشی، استفاده از روش‌های جریمه‌شده است. اگر تنها هدف در مدل‌بندی رگرسیونی برآورد و تفسیر ضرایب رگرسیونی باشد برآوردگر $\hat{\beta}$، ابزار مفیدی است اما اگر برخی ضرایب معنی‌دار نباشند و بخواهیم انتخاب متغیر را نیز همزمان با برآورد ضرایب انجام دهیم، برآوردگر لاسو[۱۳] (LASSO)، که اولین بار توسط تیشیبرانی[۱۴] (۱۹۹۶) مطرح شد، ابزار مناسب‌تری است. برآوردگر لاسو انقلاب بزرگی در مدل‌بندی رگرسیونی برپا کرد و پس از آن در مدل‌های رگرسیونی زیادی برای برآورد و انتخاب متغیر به طور همزمان توسعه داده شد. برای آگاهی بیشتر در مورد تاریخچه این برآوردگر و کاربردهای آن در ایران به تحقیقات آرست (۱۳۹۵) و نوروزی‌راد (۱۳۹۶) مراجعه کنید.

برآوردگر لاسو یک برآوردگر جریمه‌شده است. لذا در این فصل پس از معرفی برآوردگرهای جریمه‌شده به طور مختصر، برخی از توابع جریمه را نام برده و در ادامه فصل به طور اجمالی توضیحاتی بیشتر در خصوص برآوردگر لاسو، تاریخچه و چگونگی استفاده از آن می‌پردازیم.

در ادامه، برآوردگر LS پارامتر عرض از مبدأ را در مدل رگرسیون چندگانه می‌یابیم. برای این منظور فرض کنید تحت مفروضات (۳.۱)، مدل رگرسیون به صورت زیر است

$$\boldsymbol{y} = \beta_{\circ} I_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

واضح است که مولفه‌های بردار $\boldsymbol{y}$ را می‌توان به صورت زیر نوشت

$$y_i = \alpha + \beta_{۱}(x_{i۱} - \bar{x}_{۱}) + \beta_{۲}(x_{i۲} - \bar{x}_{۲}) + \cdots + \beta_p(x_{ip} - \bar{x}_p) + \epsilon_i$$

[۱۲] Friedman

[۱۳] Least absolute shrinkage and selection operator

[۱۴] Tibshirani

که در آن

$$\alpha = \beta_۰ + \beta_۱\bar{x}_۱ + \cdots + \beta_p\bar{x}_p, \qquad \bar{x}_j = \frac{۱}{n}\sum_{i=۱}^{n} x_{ij}. \tag{۷.۱}$$

بنابراین

$$(\boldsymbol{X} - \bar{\boldsymbol{X}}) = \begin{pmatrix} x_{۱۱} - \bar{x}_۱ & x_{۱۲} - \bar{x}_۲ & \dots & x_{۱p} - \bar{x}_p \\ x_{۲۱} - \bar{x}_۱ & x_{۲۲} - \bar{x}_۲ & \dots & x_{۲p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n۱} - \bar{x}_۱ & x_{n۲} - \bar{x}_۲ & \dots & x_{np} - \bar{x}_p \end{pmatrix} = \begin{pmatrix} (\boldsymbol{x}_۱ - \bar{\boldsymbol{x}})^\top \\ (\boldsymbol{x}_۲ - \bar{\boldsymbol{x}})^\top \\ \vdots \\ (\boldsymbol{x_n} - \bar{\boldsymbol{x}})^\top \end{pmatrix}$$

که در آن $\bar{\boldsymbol{x}} = (\bar{x}_۱, \bar{x}_۲, \dots, \bar{x}_p)^\top$ و $\boldsymbol{x_i} = (x_{i۱}, x_{i۲}, \dots, x_{ip})^\top$. لذا برآوردگر LS پارامتر $\boldsymbol{\beta}_۱$ به صورت

$$\hat{\boldsymbol{\beta}}_۱ = (\boldsymbol{x}_۱^\top \boldsymbol{x}_۱)^{-۱}\boldsymbol{x}_۱^\top \boldsymbol{y} \tag{۸.۱}$$

حاصل می‌شود. واضح است اگر در $E(\boldsymbol{y}) = \boldsymbol{\beta}_۰ + \boldsymbol{\beta}_۱\boldsymbol{x}_۱ + \cdots + \boldsymbol{\beta_p}\boldsymbol{x_p}$ قرار دهیم $\boldsymbol{x}_۱ = \bar{\boldsymbol{x}}_۱$، $\boldsymbol{x}_۲ = \bar{\boldsymbol{x}}_۲$، ...، $\boldsymbol{x}_p = \bar{\boldsymbol{x}}_p$، آن‌گاه نتیجه با (۷.۱) یکسان است. پس α همان $E(\boldsymbol{y}) = \bar{y}$ است و لذا

$$\hat{\alpha} = \bar{y}. \tag{۹.۱}$$

با استفاده از (۹.۱) و جایگذاری در رابطه (۷.۱) برآورد $\beta_۰$ به صورت زیر حاصل می‌شود

$$\hat{\beta}_۰ = \hat{\alpha} - \hat{\beta}_۱\bar{x}_۱ - \hat{\beta}_۲\bar{x}_۲ - \cdots - \hat{\beta}_p\bar{x}_p. \tag{۱۰.۱}$$

۲.۱ رگرسیون جریمه‌شده

از آن‌جایی که در حل معادله (۶.۱) می‌خواهیم میزان خطا را بر حسب $\boldsymbol{\beta}$ کمینه کنیم اگر مقادیر β ها بزرگ باشد برآوردگر حاصل، بدون در نظر گرفتن این بزرگی، اختلاف زیادی از β واقعی داشته و خطای برآورد زیاد می‌شود. یک راه‌حل غلبه بر این مشکل، جریمه کردن مقادیر بزرگ $\boldsymbol{\beta}$ است؛ لذا به جای کمینه‌کردن $S(\boldsymbol{\beta})$ مقدار $S(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta})$ را کمینه می‌کنیم که در آن $P_\lambda(\boldsymbol{\beta})$ تابعی بر حسب بزرگی $\boldsymbol{\beta}$ بوده و پارامتر $\lambda > ۰$ میزان بزرگی را مشخص می‌کند. برآوردگر حاصل از کمینه‌کردن عبارت فوق، برآوردگر جریمه‌شده نام دارد که به صورت زیر به‌دست می‌آید

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{Pen}} = \underset{\boldsymbol{\beta}\in R^p}{\operatorname{argmin}}\{S(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})\} \tag{۱۱.۱}$$

که در آن $P(\boldsymbol{\beta})$ همان تابع جریمه $P_\lambda(\boldsymbol{\beta})$ است با این تفاوت که پارامتر λ کنترل‌کننده میزان جریمه بوده، به صورت ضربی (برای سادگی در محاسبات) در تابع جریمه اعمال شده است. در ادامه برخی از انواع تابع جریمه $P(\boldsymbol{\beta})$ که شکل‌های متفاوتی دارند و در حالت‌های مختلفی به کار می‌روند را به عنوان نمونه آورده و برخی از خواص آن‌ها را بیان کرده و همچنین مرجع مناسبی جهت مطالعه بیشتر ارائه می‌دهیم.

۱.۲.۱ انواع تابع جریمه

توابع جریمه در رگرسیون جریمه‌شده انواع مختلفی دارد که با توجه به مسئله مورد بررسی و کاربرد مورد انتظار، رابطه ریاضی آن متفاوت است. اما یک خانواده کلی از توابع جریمه مربوط به رگرسیون جریمه‌شده بریج است که این توابع جریمه دارای ساختار زیر است

$$\boldsymbol{P(\beta) = \sum_{i=1}^{p} |\beta_i|^\gamma, \qquad 0 \leq \gamma.} \tag{۱۲.۱}$$

بدیهی است برآوردگر حاصل در رابطه (۱۱.۱) با استفاده از (۱۲.۱)، برآوردگر بریج نام دارد.

لازم به ذکر است در ادامه این گونه موارد مشابه کلمه برآوردگر بریج، همانند برآوردگر ریج داریم که دلیل نام‌گذاری، تنها تابع جریمه خاص آن است. آرست (۱۳۹۵) برآوردگر بریج را به‌طور کامل بررسی‌کرده است. به ازای $\gamma = 1$ و $\gamma = 2$ به ترتیب برآوردگرهای لاسو و ریج حاصل می‌شوند. از آن‌جایی که هدف از این فصل، توضیح بیشتر در مورد برآوردگر لاسو است و برآوردگر ریج نیز شباهت زیادی با آن (به دلیل ساختار رابطه (۱۲.۱)) دارد، از بررسی جزئیات سایر توابع جریمه کاسته و پس از توضیح برآوردگر ریج، برآوردگر لاسو را مورد بررسی قرار می‌دهیم.

- برآوردگر ریج: با انتخاب

$$P(\boldsymbol{\beta}) = \sum_{i=1}^{p} |\beta_i|^2$$

برآوردگر حاصل در رابطه (۱۱.۱)، برآوردگر ریج نام دارد. در این حالت

$$\begin{aligned} S(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X\beta})^\top (\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{\lambda \beta^\top \beta} \\ &= \boldsymbol{y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X \beta + \lambda \beta^\top \beta} \end{aligned}$$

با مشتق‌گیری نسبت به $\boldsymbol{\beta}$ و مساوی صفر قرار دادن حاصل آن داریم

$$\boldsymbol{2X^\top X\beta + 2\lambda\beta = 2X^\top y}$$

و در نتیجه برآوردگر ریج با حل رابطه زیر نسبت به $\boldsymbol{\beta}$

$$(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\lambda I_p})\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y}$$

به صورت

$$\hat{\boldsymbol{\beta}}^R(\lambda) = (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\lambda I_p})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

حاصل می‌شود. برآوردگر ریج اولین بار توسط هورل و کنارد[۱۵] (۱۹۷۰) ارائه شده است. معمولا از برآوردگر ریج زمانی که هم‌خطی بین ستون‌های ماتریس $\boldsymbol{X}^\top \boldsymbol{X}$ وجود دارد، استفاده می‌شود. برای توضیح بیشتر و جلوگیری از ارائه مطالب تکراری، منبع نیرومند (۱۳۸۷) را ببینید. پایان‌نامه‌های زیادی در رابطه با برآوردگر ریج، خواص و کاربردهای آن نوشته شده‌اند که از جمله آن‌ها می‌توان به نجاریان (۱۳۹۰)، برزویی بیدگلی (۱۳۹۳)، و آرست (۱۳۹۵) اشاره کرد. در ادامه به چند خاصیت مهم برآوردگر ریج اشاره می‌کنیم.

امید ریاضی و واریانس برآوردگر ریج $\hat{\boldsymbol{\beta}}^R(\lambda)$ به ترتیب عبارتند از

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}^R(\lambda)] &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta}, \\ Var[\hat{\boldsymbol{\beta}}^R(\lambda)] &= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}. \end{aligned}$$

قضیه ۱.۲.۱. تحت مفروضات مدل (۳.۱)، در صورتی که $\boldsymbol{X}^\top \boldsymbol{X}$ معکوس‌پذیر باشد، برآوردگر ریج، تحت نرم L_q، $q > 0$ یک برآوردگر انقباضی است. به عبارتی داریم

$$\|\hat{\boldsymbol{\beta}}^R(\lambda)\|_q^q \leq \|\hat{\boldsymbol{\beta}}\|_q^q. \qquad (۱۳.۱)$$

برهان. از آن جایی که $\boldsymbol{X}^\top \boldsymbol{X}$ معکوس‌پذیر است، می‌توان نوشت

$$\begin{aligned} \hat{\boldsymbol{\beta}}^R(\lambda) &= (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\lambda I_p})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ &= (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\lambda I_p})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) \hat{\boldsymbol{\beta}} \\ &= (\boldsymbol{I_p} + \boldsymbol{\lambda} (\boldsymbol{X}^\top \boldsymbol{X})^{-1})^{-1} \hat{\boldsymbol{\beta}}. \qquad (۱۴.۱) \end{aligned}$$

حال تجزیه طیفی[۱۶] $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Gamma}^\top$ را به صورت $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ در نظر بگیرید که در آن $\boldsymbol{\Gamma}$ یک ماتریس متعامد در اندازه $p \times p$ شامل بردارهای ویژه و $\boldsymbol{\Lambda}$ ماتریس قطری شامل مقادیر ویژه $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ است. در این صورت، نتیجه می‌شود

$$\|\hat{\boldsymbol{\beta}}^R(\lambda)\|_q^q = \|(\boldsymbol{I_p} + \lambda \boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Gamma}^\top)^{-1} \hat{\boldsymbol{\beta}}\|_q^q$$

[۱۵] Hoerl and Kennard

[۱۶] Spectral decomposition

$$\begin{aligned} &= \|(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top + \lambda\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top)^{-1}\hat{\boldsymbol{\beta}}\|_q^q \\ &= \|\boldsymbol{\Gamma}(\boldsymbol{I_p} + \lambda\boldsymbol{\Lambda})^{-1}\boldsymbol{\Gamma}^\top\hat{\boldsymbol{\beta}}\|_q^q \\ &= \left(\|\boldsymbol{\Gamma}(\boldsymbol{I_p} + \lambda\boldsymbol{\Lambda})^{-1}\boldsymbol{\Gamma}^\top\hat{\boldsymbol{\beta}}\|_q^2\right)^{\frac{q}{2}} \\ &\leq \left(\|\boldsymbol{\Gamma}^\top\hat{\boldsymbol{\beta}}\|_q^2\right)^{\frac{q}{2}} \\ &= \|\hat{\boldsymbol{\beta}}\|_q^q \end{aligned}$$

□

در ادامه نشان می‌دهیم برآوردگر LS در یک مسئله رگرسیون با مشاهدات تجمیع شده[۱۷]، برآوردگر ریج را نتیجه می‌دهد.

برای این منظور ماتریس گسترش‌یافته $\boldsymbol{X_\lambda}$ و بردار $\boldsymbol{y_\lambda}$ به صورت زیر را در نظر بگیرید

$$\boldsymbol{X_\lambda} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \\ \sqrt{\lambda} & \circ & \dots & \circ \\ \circ & \sqrt{\lambda} & \dots & \circ \\ \vdots & \vdots & \cdots & \vdots \\ \circ & \circ & \dots & \sqrt{\lambda} \end{pmatrix}, \qquad \boldsymbol{y_\lambda} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ \circ \\ \circ \\ \vdots \\ \circ \end{pmatrix}.$$

واضح است که

$$\boldsymbol{X_\lambda} = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda}\boldsymbol{I_p} \end{pmatrix}, \qquad \boldsymbol{y_\lambda} = \begin{pmatrix} \boldsymbol{y} \\ \circ \end{pmatrix}.$$

در این حالت برآوردگر LS بردار پارامتر $\boldsymbol{\beta}$ در مدل

$$\boldsymbol{y}_\lambda = \boldsymbol{X}_\lambda\boldsymbol{\beta} + \boldsymbol{\epsilon}_\lambda$$

(که در آن $\boldsymbol{\epsilon}_\lambda$ به طور مشابه ساخته می‌شود) به صورت زیر محاسبه می‌گردد

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\operatorname{argmin}}(\boldsymbol{y}_\lambda - \boldsymbol{X}_\lambda\boldsymbol{\beta})^\top(\boldsymbol{y}_\lambda - \boldsymbol{X}_\lambda\boldsymbol{\beta}) \\ &= (\boldsymbol{X}_\lambda^\top\boldsymbol{X})^{-1}\boldsymbol{X}_\lambda^\top\boldsymbol{y}_\lambda \end{aligned}$$

[۱۷] Augmented

$$\begin{aligned} &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ &= \hat{\boldsymbol{\beta}}^R(\lambda). \end{aligned} \tag{۱۵.۱}$$

در انتها، ساختار برآوردگر ریج را در حالتی که ماتریس $\boldsymbol{X}$ متعامد است بررسی می‌کنیم. در این حالت، $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_p$ و برآوردگر LS به صورت

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ &= \boldsymbol{X}^\top \boldsymbol{y} \end{aligned}$$

است. بنابراین

$$\begin{aligned} \hat{\boldsymbol{\beta}}^R(\lambda) &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ &= (\boldsymbol{I}_p + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ &= \frac{1}{(\lambda+1)} \boldsymbol{X}^\top \boldsymbol{y} \\ &= \frac{1}{(\lambda+1)} \hat{\boldsymbol{\beta}} \end{aligned}$$

لذا همواره

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}^R(\lambda)) &= \frac{1}{(\lambda+1)^2} Var(\hat{\boldsymbol{\beta}}) \\ &\leq Var(\hat{\boldsymbol{\beta}}) \end{aligned} \tag{۱۶.۱}$$

و این نشان می‌دهد در حالت متعامد همواره برآوردگر ریج واریانس کمتری نسبت به برآوردگر LS دارد.

- برآوردگر لاسو: تیبشیرانی (۱۹۹۶) با در نظر گرفتن تابع جریمه $P(\boldsymbol{\beta}) = \sum_{i=1}^{p} |\beta_i|$ برآوردگر زیر را پیشنهاد کرد

$$\hat{\boldsymbol{\beta}}^{LASSO} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{\boldsymbol{p}}}{\operatorname{argmin}} \left\{ S(\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i| \right\}. \tag{۱۷.۱}$$

برآوردگر (۱۷.۱)، برآوردگر لاسو نام دارد. اگر چه برآوردگر لاسو صورت بسته‌ای، همانند برآوردگر ریج، ندارد ولی می‌تواند علاوه بر برآورد ضرایب $\boldsymbol{\beta}$، انتخاب متغیر نیز انجام دهد. در این حالت ضرایب $\hat{\boldsymbol{\beta}}$ به سمت صفر منقبض شده و برخی از آن‌ها مقدار صفر را انتخاب می‌کنند. توضیحات بیشتر در خصوص این برآوردگر در بخش ۳ آورده شده است.

- برآوردگر لاسوی تطبیقی: ژو[۱۸](۲۰۰۶) پیشنهاد کرد به جای تابع جریمه قدر مطلق به تنهایی، از تابع جریمه

$$P(\boldsymbol{\beta}) = \hat{\boldsymbol{W}}|\boldsymbol{\beta}| = \sum_{j=1}^{p} \hat{W}_j|\beta_j|$$

استفاده شود که در آن

$$\hat{\boldsymbol{W}} = \text{diag}(\hat{W}_1, \ldots, \hat{W}_p) \quad , \quad |\boldsymbol{\beta}| = (|\beta_1|, \ldots, |\beta_p|)^\top$$

و

$$\hat{W}_j = \frac{1}{|\hat{\beta}_j|^r}, \qquad r > 0$$

که $\hat{\beta}_j$ یک برآوردگر $\sqrt{n}$ـسازگار برای $\boldsymbol{\beta}_j$ می‌باشد. می‌توان از برآوردگرهای LS، لاسو و ریج به عنوان $\hat{\beta}_j$ ، $j = 1, \ldots, p$ استفاده کرد. این برآوردگر دارای خاصیت اوراکل[۱۹] (معجز)، در زیر، است در صورتی که لاسو این خاصیت را ندارد.

- خاصیت ۱: ضرایب غیر صفر را به درستی مشخص می‌کند. (یعنی زیر مدل درست را تعیین می‌کند.)
- خاصیت ۲: دارای نرخ برآورد بهینه است. به عبارتی ماتریس‌کوواریانس آن، در حالت حدی $(n \to \infty)$، با ماتریس‌کوواریانس برآوردگر زیر مدل برابر است.

برآوردگر لاسوی تطبیقی به صورت زیر است

$$\hat{\boldsymbol{\beta}}^{ALASSO} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{\boldsymbol{p}}}{\text{argmin}} \left\{ S(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \hat{W}_j|\beta_j| \right\}.$$

- لاسوی ذوب‌شده[۲۰] : در بعضی مواقع ممکن است بتوان یک ترتیب معنی‌دار بین متغیرها در نظر گرفت. جریمه لاسوی ذوب‌شده هم نرم L_1 ضرایب و هم تفاضل‌های پشت سرهم آن را جریمه می‌کند. از کاربرد این روش می‌توان در بحث‌های سری زمانی یا داده‌های از نوع تصویر اشاره کرد. تیبشیرانی و همکاران

[۱۸] Zou
[۱۹] Oracle
[۲۰] Fused LASSO

(۲۰۰۵) جریمه لاسوی ذوب‌شده را به صورت

$$P_\lambda(\boldsymbol{\beta}) = \lambda_1 \sum_{i=1}^{p} |\boldsymbol{\beta}_i| + \lambda_2 \sum_{i=2}^{p} |\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i-1}| \qquad (۱۸.۱)$$

معرفی کردند. برآوردگر لاسوی ذوب‌شده به صورت زیر است:

$$\hat{\boldsymbol{\beta}}^{Fused} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ S(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) \right\},$$

که $P_\lambda(\boldsymbol{\beta})$ در (۱۸.۱) آمده است.

- برآوردگر اسکد[۲۱] : در سال ۲۰۰۱، فن و لی[۲۲]، سه شرط برای یک تابع جریمه خوب ارائه کردند

  - شرط ۱ (نااریبی[۲۳]): برای جلوگیری از اریبی مدل، باید پارامتر برآوردشده به مقادیر بزرگ پارامتر نزدیک باشد.
  - شرط ۲ (تنکی [۲۴]): زمانی که ضرایب کوچک هستند، برآوردگر آن‌ها را صفر می‌کند تا پیچیدگی مدل را کاهش دهد.
  - شرط ۳ (پیوستگی[۲۵]): برآوردگر بدست‌آمده در داده‌ها پیوسته است تا از ناپایداری مدل جلوگیری کند.

همچنین آن‌ها بیان کردند که توابع جریمه محدب این سه شرط را دارا نیستند و اگر مقعر باشد آن‌ها را داراست. برای این منظور تابع جریمه اسکد که مقعر است را معرفی کردند.

فن و لی (۲۰۰۱) تابع جریمه اسکد را به صورت زیر ارائه کردند

$$P_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j| & |\beta_j| \le \lambda \\ -\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)} & \lambda < |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & |\beta_j| > a\lambda \end{cases}.$$

در عمل مقدار بهینه $a = 3.7$ استفاده می‌شود. تابع جریمه اسکد به طور پیوسته روی $\mathbb{R} - \{0\}$ مشتق‌پذیر است ولی در صفر مشتق‌پذیر نمی‌باشد و دارای شروط

[۲۱] The Smoothly clipped absolute deviation
[۲۲] Fan and Li
[۲۳] Unbiasedness
[۲۴] Sparsity
[۲۵] Continuity

سه‌گانه تابع جریمه خوب است. بنابراین برآوردگر اسکد به صورت زیر بدست می‌آید

$$\hat{\boldsymbol{\beta}}^{SCAD} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\operatorname{argmin}} \left\{ S(\boldsymbol{\beta}) + \sum_{j=1}^{p} P_{\lambda}(\beta_j) \right\}.$$

- برآوردگر الاستیک‌نت[۲۶]: هرگاه تابع جریمه به صورت

$$P(\boldsymbol{\beta}) = \lambda_1 \sum_{i=1}^{n} |\boldsymbol{\beta}_i| + \lambda_2 \sum_{i=1}^{n} \boldsymbol{\beta}_i^2$$

در نظر گرفته شود به آن تابع جریمه الاستیک‌نت گویند که λ_1 و λ_2 مقادیر نامنفی اختیار می‌کنند. اگر λ_1 مقدار صفر را اختیار کند، جریمه ریج و اگر مقدار λ_2 را صفر قرار دهیم جریمه لاسو حاصل می‌شود پس جریمه الاستیک‌نت ترکیبی از جریمه های ریج و لاسو است. برآوردگر الاستیک‌نت به صورت زیر به‌دست می‌آید

$$\hat{\boldsymbol{\beta}}^{Enet} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\operatorname{argmin}} \left\{ S(\boldsymbol{\beta}) + \lambda_1 \sum_{i=1}^{q} |\boldsymbol{\beta}_i| + \lambda_2 \sum_{i=1}^{q} \boldsymbol{\beta}_i^2 \right\}. \qquad (۱۹.۱)$$

این برآوردگر توسط ژو و هستی[۲۷] (۲۰۰۵) ارائه شد. این برآوردگر برخی از ضرایب را صفر می‌کند و برای متغیرهای همبسته اثر یکسانی در نظر می‌گیرد. بنابراین این روش زمانی که گروه‌بندی متغیرهای مستقل زیاد است مورد استفاده قرار می‌گیرد.

- برآوردگر لاسوی گروهی[۲۸]: یوان و لین[۲۹] (۲۰۰۶) لاسو گروهی را معرفی کردند. این برآوردگر اجازه می‌دهد گروهی از متغیرهای همراه هم از مدل حذف شوند. به عبارت ساده‌تر برخلاف لاسو که انتخاب متغیر تکی صورت می‌گیرد در این حالت انتخاب متغیر گروهی امکان‌پذیر است. به عنوان مثال در متغیرهای رسته‌ای[۳۰] که سطوح به صورت گروهی از متغیرهای همراه دودویی کدبندی می‌شود، می‌توان از لاسوی گروهی برای انتخاب سطوح معنی‌دار به طور جمعی استفاده کرد. در این حالت تابع جریمه به صورت

$$P(\boldsymbol{\beta}) = \sum_{i=1}^{j} \|\boldsymbol{\beta}_i\|_{\boldsymbol{K}_i}$$

---

[۲۶] Elastic Net
[۲۷] Zou and Hastie
[۲۸] Group lasso
[۲۹] Yuan and Lin
[۳۰] Categorical variables

است که j تعداد گروه‌ها یا رسته‌ها است و

$$\|\boldsymbol{\beta}_i\|_{\boldsymbol{K}_i} = (\boldsymbol{\beta}_i^\top \boldsymbol{K}_i \boldsymbol{\beta}_i)^{\frac{1}{2}}$$

و $\boldsymbol{K}_i$ یک ماتریس معین مثبت است. برآوردگر لاسوی گروهی به صورت زیر است

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \sum_{i=1}^{p} \boldsymbol{X}_i\boldsymbol{\beta}_i\|_2^2 + \lambda \sum_{i=1}^{p} \|\boldsymbol{\beta}_i\|_{\boldsymbol{K}_i} \right\} \qquad (۲۰.۱)$$

۳.۱ برآوردگر لاسو

اگرچه که لاسو، در ابتدا برای مسائل رگرسیون خطی معرفی شد، این رهیافت در سری‌های زمانی توسط نویسندگان بسیاری به ویژه در حالت مدل اتورگرسیو به کار گرفته شد (وانگ[۳۱] و همکاران ۲۰۰۷، ناردی و رینالدو[۳۲]، ۲۰۱۱). از انواع دیگر لاسو می‌توان به لاسوی تطبیقی[۳۳] (ژو[۳۴]، ۲۰۰۶) و جزئیات آن (بولمن و ون دگیر[۳۵]، ۲۰۱۱) اشاره کرد. این روش به عنوان مجموعه‌ای از تکنیک‌هایی برای مدل بعد بالا به کار گرفته می‌شود. واریان[۳۶] (۲۰۱۴) و فن و همکاران[۳۷] (۲۰۱۱) به مروری بر این روش‌ها می‌پردازند. هر دوی این مقاله‌ها، تکنیک‌های بعد بالا را خلاصه می‌کند و مثال‌های کاربردی در آن‌ها ارائه شده است. همچنین مثال کاربردی را می‌توان در بای و ان جی[۳۸] (۲۰۰۸) مشاهده کرد. نویسندگان این مقاله، کارایی چندین روش داده‌های بعد بالا را برای پیش‌بینی سری‌های زمانی مقایسه کردند.

اگر چه رگرسیون ریج، بهبود پیش‌گویی را نتیجه می‌دهد ولی با توجه به حضور همه متغیرها در آن، تفسیرپذیری به‌سادگی امکان‌پذیر نیست. از طرف دیگر، حضور متغیرهایی که با متغیر پاسخ در یک مدل ارتباط ندارند، در فرآیند برآورد، منجر به برآوردگری با واریانس بزرگ می‌شود.

روشی ساده و بدیهی برای حل این مسئله، انتخاب یک زیرمجموعه از متغیرهای توضیحی بااهمیت و تاثیرگذار در مدل می‌باشد. انتخاب این متغیرها می‌تواند بر اساس معنی‌داری آن‌ها در مدل یا برخی از معیارهای اطلاع مانند AIC [۳۹] (آکائیک، ۱۹۷۳)،

[۳۱] Wang
[۳۲] Nardi and Rinaldo
[۳۳] Adaptive lasso
[۳۴] Zhou
[۳۵] Buhlmann and Van de Geer
[۳۶] Varian
[۳۷] Fan et al.
[۳۸] Bay and NG
[۳۹] Akaike Information Criterion

BIC [۴۰] (شوارتز[۴۱] ، ۱۹۷۸)، و C_pـ مالو [۴۲] (مالوز، ۱۹۷۳) انجام شود. اما مشکل بزرگی که در هنگام استفاده از این معیارها مطرح می‌شود تعداد کل زیرمجموعه‌هایی از متغیرها است که باید در نظر گرفته شوند، که به‌طور نمایی با افزایش p رشد می‌کند و از نظر محاسباتی برای تعداد متغیرهای بیشتر از ۴۰ یا ۵۰، تقریبا نشدنی است. مشکل دیگر این روش‌ها، ویژگی گسسته بودن آن‌هاست. به‌عبارت دیگر، تغییر بسیار کوچکی در داده‌ها، می‌تواند برآوردهای کاملا متفاوتی را نتیجه دهد. در مجموع، روش‌های انتخاب بهترین زیرمجموعه، اغلب ناپایدار و به شدت تغییرپذیر هستند. روش دیگر، برای رویارویی با این مسئله، استفاده از رگرسیون جریمه‌شده است. اگر چه به‌کاربردن جریمه ریج، در رگرسیون جریمه‌شده، به دستیابی برآوردگرهایی با واریانس کمتر کمک بسیاری می‌کند اما وقتی مقدار قدرمطلق ضرایب رگرسیونی بزرگ باشد، برآوردگر ریج اریبی زیادی به سمت صفر ایجاد می‌کند و از طرف دیگر، با وجودی که متغیرهای بی‌اهمیت به‌سمت صفر منقبض شده‌اند، هم‌چنان در مدل حضور دارند که باعث می‌شود تفسیرپذیری مدل به‌راحتی انجام نشود. در این راستا، تیبشیرانی (۱۹۹۶) روشی جدید برای انتخاب متغیرها پیشنهاد نمود که منجر به مدل‌های دقیق، پایا و صرفه‌جو (تفسیرپذیرتر) می‌شود و نام آن را لاسو نامید.

از دیدگاه آماری، لاسو مخفف عملگر انتخاب‌کننده، منقبض‌کننده و کمینه‌کننده از جنس قدرمطلق است. برآوردگر لاسو، نوع جریمه‌شده برآوردگر LS است که بر اساس ویژگی تنکی نرم L_1، در سال‌های اخیر مورد توجه قرار گرفته است. برآوردگر لاسو از حل مسئله بهینه‌سازی زیر به‌دست می‌آید:

$$\hat{\boldsymbol{\beta}}_n^{LASSO} = \min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 \right\} \quad \text{s.t.} \quad \sum_{i=1}^{p} |\beta_j| \leq t, \qquad (۲۱.۱)$$

که در آن t یک مقدار ثابت است. اگر $t = 0$، آن‌گاه متغیری در مدل وجود نخواهد داشت در حالی‌که اگر $t = \infty$، آن‌گاه مدل، کامل خواهد بود. فرض کنید $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_1^*, \ldots, \hat{\beta}_p^*)^\top$ یک برآوردگر اولیه برای $\boldsymbol{\beta}$ است که معمولا این برآوردگر را $\hat{\boldsymbol{\beta}}^{LS}$ در نظر می‌گیرند. اگر $t > \sum_{i=1}^{p} |\hat{\beta}_i^*|$، روش لاسو منجر به برآوردگر LS می‌شود.

اگر $0 < t < \sum_{j=1}^{p} |\hat{\beta}_j^{LS}|$ باشد، آن‌گاه مسئله بهینه‌سازی (۲۱.۱) معادل با

$$\hat{\boldsymbol{\beta}}^{LASSO} = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \qquad (۲۲.۱)$$

[۴۰] Bayesian Information Criterion

[۴۱] Schwarz

[۴۲] Mallows

است، که در آن λ پارامتر تنظیم‌کننده است و سطح تنکی (تعداد پارامترهای صفر) را در برآوردگر لاسو کنترل می‌کند. وقتی $\lambda \to 0$، $\hat{\boldsymbol{\beta}}^{LASSO} \to \hat{\boldsymbol{\beta}}^{LS}$ و وقتی $\lambda \to \infty$، $\hat{\boldsymbol{\beta}}^{LASSO} \to \mathbf{0}$. به‌عبارت دیگر، هرچه جریمه بزرگتری به‌کار برده شود، تعداد بیشتری از ضرایب به‌سمت صفر منقبض می‌شوند. در حقیقت λ و t رفتار یا رابطه‌ای معکوس با یکدیگر دارند.

به رگرسیون لاسو، به خاطر شکل تابع جریمه آن، رگرسیون جریمه‌شده L_1 نیز می‌گویند. انتخاب این تابع جریمه باعث می‌شود که برآوردگر لاسو، ضرایب را به سمت صفر منقبض کرده و متغیرهای اضافی را از مدل حذف نماید. در واقع، برآوردگر لاسو، همزمان هم انتخاب متغیر انجام می‌دهد و هم ضرایب را منقبض می‌کند.

تابع جریمه قدرمطلق باعث می‌شود که جواب صریحی برای برآوردگر لاسو وجود نداشته باشد. تیبشیرانی (۱۹۹۶) این برآوردگر را از طریق برنامه‌ریزی درجه دوم به‌دست آورده بود. افرون و همکاران (۲۰۰۴) نوع دیگری از رگرسیون گام به گام را به‌نام رگرسیون کمترین زاویه[۴۳] (LAR) معرفی نمودند که برآوردهای لاسویی را نتیجه می‌دهد که هزینه محاسباتی آن با برآوردهای کمترین توان‌های دوم برابر است.

فن و لی (۲۰۰۱) نشان دادند که یک تابع جریمه خوب باید برآوردگری با سه ویژگی مطلوب نااریبی، تُنُکی و پیوستگی را نتیجه دهد. جریمه L_1 برآوردگرهایی را نتیجه می‌دهد که تُنُک، نسبت به داده‌ها پیوسته و اریب هستند زیرا اندازه انقباضی که به ضرایب رگرسیونی کوچک و بزرگ اختصاص می‌دهد، یکسان است (لازم به ذکر است این مورد در خصوص جریمه L_2 صادق نیست). هم‌چنین، لاسو برای یک مدل رگرسیونی خطی با p متغیر توضیحی و n مشاهده، حداکثر n متغیر انتخاب می‌کند و بنابراین اگر متغیرهای بیشتری (بیشتر از n) در مدل معنی‌دار باشند، توسط لاسو انتخاب نمی‌شوند. لاسو، نمی‌تواند متغیرهایی که اثر معنی‌داری کمی دارند را تشخیص دهد و آن را از مدل حذف می‌کند. برای داده‌های هم‌خط مناسب نیست زیرا در بین گروه متغیرهایی که وابستگی شدیدی دارند، تنها یکی از آن‌ها را انتخاب می‌کند. اگر همبستگی بالایی بین متغیرهای توضیحی وجود داشته باشد، کارآیی پیش‌گویی لاسو به خوبی برآوردگر ریج نیست. بزرگترین عیب لاسو این است که خاصیت اوراکل را ندارد.

(۱) زیر مدل درست را مشخص کند به‌عبارتی، مجموعه $\{j : \hat{\beta}_j \neq 0\}$ را طوری مشخص کند که $\mathcal{A} = \{j : \hat{\beta}_j \neq 0\}$، که در آن A زیر مجموعه پارامترهای مدل درست است.

[۴۳] Least Angle Regression

(۲) دارای نرخ برآورد بهینه باشد به‌عبارتی،

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\delta) - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Sigma^*)$$

که در آن Σ^* ماتریس کوورایانس زیرمدل درست است.

۱.۳.۱ برآوردگر لاسو در حالت متعامد

برآوردگر لاسو دارای صورت بسته نیست و نمی‌توان جواب صریحی به صورت تحلیلی، در حالت کلی، برای آن پیدا کرد. اما در حالت متعامد می‌توان رابطه تحلیلی $\hat{\boldsymbol{\beta}}^{LASSO}$ را ارائه کرد که در قالب لم زیر آورده شده است.

لم ۱.۳.۱. تحت مفروضات مدل (۳.۱) فرض کنید $\boldsymbol{X}$ یک ماتریس متعامد باشد. در این صورت

$$\hat{\boldsymbol{\beta}}^{LASSO} = (\hat{\beta}_i^{LASSO}, \qquad i = ۱, \dots, p)^\top$$

که در آن

$$\hat{\beta}_i^{LASSO} = \text{sign}(\hat{\beta}_i^{LS})\left(|\hat{\beta}_i^{LS}| - \frac{\lambda}{۲}\right)^+, \qquad i = ۱, \dots, p. \tag{۲۳.۱}$$

برهان. طبق فرض داریم $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}_p$ و $(\boldsymbol{X}^\top\boldsymbol{X})^{-۱} = \boldsymbol{I}_p$. لذا $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-۱}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{X}^\top\boldsymbol{y}$ پس می‌توان نوشت

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{LASSO} &= \min_{\beta\in\mathbb{R}^p}\left\{S(\boldsymbol{\beta}) + \lambda\sum_{i=۱}^{p}|\beta_i|\right\} \\
&= \min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \lambda\sum_{i=۱}^{p}|\beta_i|\right\} \\
&\propto \min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{-[\hat{\boldsymbol{\beta}}^{\boldsymbol{LS}}]^\top\boldsymbol{\beta} - \boldsymbol{\beta}^\top\hat{\boldsymbol{\beta}}^{\boldsymbol{LS}} + \boldsymbol{\beta}^\top\boldsymbol{\beta} + \lambda\sum_{i=۱}^{p}|\beta_i|\right\} \\
&= \min_{\beta_۱,\dots,\beta_p}\sum_{i=۱}^{p}(-۲\hat{\beta}_i^{LS}\beta_i + \beta_i^۲ + \lambda|\beta_i|) \\
&= \sum_{i=۱}^{p}\min_{\boldsymbol{\beta}_i\in\mathbb{R}^p}\left\{-۲\hat{\beta}_i^{LS}\beta_i + \beta_i^۲ + \lambda|\beta_i|\right\}.
\end{aligned}$$

حال ضرایب رگرسیونی را کمینه می‌کنیم برای این منظور داریم

(۲۴.۱)

$$\min_{\boldsymbol{\beta}_i\in\mathbb{R}^p}\left\{-۲\hat{\beta}_i^{LS}\beta_i + \beta_i^۲ + \lambda|\beta_i|\right\} = \begin{cases}\min_{\boldsymbol{\beta}_i\in\mathbb{R}^p}\left\{-۲\hat{\beta}_i^{LS}\beta_i + \beta_i^۲ + \lambda\beta_i\right\} & \text{اگر} \quad \beta_i > ۰ \\ \min_{\boldsymbol{\beta}_i\in\mathbb{R}^p}\left\{-۲\hat{\beta}_i^{LS}\beta_i + \beta_i^۲ - \lambda\beta_i\right\} & \text{اگر} \quad \beta_i < ۰\end{cases}$$

حاصل سمت راست عبارت (۲۴.۱) به صورت زیر است

$$\hat{\beta}_i^{LASSO}(\lambda) = \begin{cases} \hat{\beta}_i^{LS} - \frac{1}{2}\lambda & \text{اگر} \quad \beta_i > 0 \\ \hat{\beta}_i^{LS} + \frac{1}{2}\lambda & \text{اگر} \quad \beta_i < 0 \end{cases}$$

و اثبات کامل است. □

فصل ۲

برآوردگر لاسوی مشتق‌پذیر

۱.۲ مقدمه

لازم به ذکر است که تابع جریمه لاسو در نقطه صفر مشتق‌پذیر نبوده و این باعث می‌شود خواص بهینه سه‌گانه یک تابع جریمه را، دارا نباشد. به منظور غلبه بر این مشکل و بهبود تابع جریمه لاسو، حاصلی‌مشهدی و وینچوتی[۱] (۲۰۱۶) تابع جریمه لاسوی مشتق‌پذیر را معرفی کردند. در این فصل نتایج آن‌ها را مورد بررسی قرار داده و پس از مطالعه خواص برآوردگر لاسوی مشتق‌پذیر، در یک مثال شبیه‌سازی رفتار این برآوردگر را در مقایسه با سایر برآوردگرهای رقیب تحلیل می‌کنیم.

۲.۲ تابع جریمه مشتق‌پذیر

برای این که بتوان تابع جریمه لاسوی مشتق‌پذیر را تعریف کرد، ابتدا باید این نکته را بررسی کنیم که در چه حالتی می‌توانیم مشتق‌پذیری تابع قدر مطلق را مورد مطالعه قرار دهیم. بدیهی است تابع قدر مطلق به خودی خود در نقطه صفر مشتق‌پذیر نبوده اما بعضی تقریب‌های آن، در صفر مشتق‌پذیر هستند.

[۱] Haselimashhadi and Vinciotti

اشمیت [۲] و همکاران (۲۰۰۷) تقریب زیر را برای تابع $|x|$ به منظور افزایش قدرت بهینه‌سازی پیشنهاد کردند که در نقطه صفر مشتق‌پذیر است:

$$|x| \approx s \log(۲ + e^{-\frac{x}{s}} + e^{\frac{x}{s}}), \qquad s \in \mathbb{R}^+.$$

پس از آن رامیرز و سانچزا[۳] (۲۰۱۴) بهترین تقریب هموار را برای تابع $|x|$ به صورت زیر ارائه کردند

$$|x| \approx \sqrt{x^۲ + s}, \qquad s \in \mathbb{R}^+$$

با تمام تلاش‌های فوق، در نهایت حاصلی‌مشهدی و وینچوتی (۲۰۱۶)، پیشنهاد کردند به عنوان تقریبی هموار از تابع $|x|$ می‌توان از تابع خطا (تعریف آ.۲.۰ را در پیوست آ ببینید) استفاده نمود. دلیل استفاده از این تقریب سادگی آن در محاسبات بعدی نسبت به دو تقریب فوق است. بنابراین آن‌ها تابع جریمه لاسوی مشتق‌پذیر را به صورت زیر پیشنهاد نمودند:

$$P(\boldsymbol{\beta}) = \sum_{i=۱}^{p} P(\boldsymbol{\beta}_i, s) \tag{۱.۲}$$

که در آن

$$P(\boldsymbol{\beta}_i, s) = \boldsymbol{\beta}_i \left(\frac{۲}{\sqrt{\pi}} \int_{۰}^{\frac{\boldsymbol{\beta}_i}{s}} \exp\{-t\}^۲ dt \right), \qquad s \in \mathbb{R}^+. \tag{۲.۲}$$

در ادامه تابع زیر را در نظر بگیرید

$$P(x, s) = x \left(\frac{۲}{\sqrt{\pi}} \int_{۰}^{\frac{x}{s}} \exp\{-t\}^۲ dt \right), \qquad s \in \mathbb{R}^+. \tag{۳.۲}$$

این تابع جریمه را می‌توان بر حسب تابع خطا به صورت‌های مختلف زیر نوشت

$$\begin{aligned} P(x, s) &= x \ \text{erf}(\frac{x}{s}) \\ &= x \left(۱ - \text{erf}^{\text{c}} \left(\frac{x}{s} \right) \right) \\ &= x \left(۲ \Phi \left(\frac{x}{s}, ۰, \frac{۱}{\sqrt{۲}} \right) - ۱ \right), \end{aligned}$$

که در آن $\Phi(x, \mu, \sigma)$ تابع توزیع تجمعی نرمال در نقطه x با میانگین μ و انحراف معیار σ است. ویژگی‌های این تابع را در قالب قضیه زیر بیان می‌کنیم.

[۲] Schmidt

[۳] Ramirez and Sanchez

قضیه ۱.۲.۲. تابع $P(x,s)$ در رابطه (۳.۲) دارای ویژگی‌های زیر است

(۱) به ازای هر s، $P(\circ,s)=\circ$.

(۲) تابع $P(x,s)$، نسبت به x دو بار مشتق‌پذیر بوده که مشتقات آن به صورت زیر است

$$\begin{aligned}\frac{d}{dx}P(x,s)&=erf\left(\frac{x}{s}\right)+۲\phi\left(\frac{x}{s},\circ,\frac{۱}{\sqrt{۲}}\right)\frac{x}{s}\\ \frac{d^۲}{dx}P(x,s)&=\frac{۲}{s}\phi\left(\frac{x}{s},\circ,\frac{۱}{\sqrt{۲}}\right)+\frac{۲}{s}\phi\left(\frac{x}{s},\circ,\frac{۱}{\sqrt{۲}}\right)-\frac{۴}{x}\left(\frac{x}{s}\right)^۳\phi\left(\frac{x}{s},\circ,\frac{۱}{\sqrt{۲}}\right)\\ &=۴\phi\left(\frac{x}{s},\circ,\frac{۱}{\sqrt{۲}}\right)\frac{۱}{s}\left(۱-\left(\frac{x}{s}\right)^۲\right),\end{aligned}$$

که در آن $\phi(x,\mu,\sigma)$ تابع چگالی توزیع نرمال در نقطه x با میانگین μ و انحراف معیار σ است.

(۳) تابع $P(x,s)$ نسبت به x محدب نیست.

(۴) وقتی $s\to\circ$، تابع با یک نرخ نمایی به $|x|$ به شدت همگرا می‌شود. در حقیقت،

$$||x|-p(x,s)|\le ۲s\phi\left(\frac{x}{s},\circ,\frac{۱}{\sqrt{۲}}\right)\qquad\forall x,s>\circ\qquad(۴.۲)$$

(۵) اگر $s=\frac{۲}{\sqrt{\pi}}$، تابع جریمه رفتاری شبیه تابع جریمه ریج در همسایگی صفر دارد. برای xهای کوچک داریم

$$\begin{aligned}\frac{۲}{\sqrt{\pi}}\int_\circ^{\frac{x}{s}}\exp\{-t^۲\}dt&\approx\frac{۲}{\sqrt{\pi}}\frac{x}{s}\exp\left\{-\left(\frac{x}{s}\right)^۲\right\}\\ &=x^۲\frac{۲}{x}\phi\left(\frac{x}{s},\circ,\frac{۱}{\sqrt{۲}}\right).\end{aligned}$$

برهان. ۱. برای اثبات (۱)، به راحتی مشاهده می‌شود

$$P(\circ,s)=\circ\times\left(\frac{۲}{\sqrt{\pi}}\int_\circ^{\frac{\circ}{s}}\exp\{-t\}^۲dt\right)=\circ$$

۲. مشتق اول نسبت به x به صورت زیر بدست می‌آید

$$\frac{d}{dx}P(x,s)=\frac{۲}{\sqrt{\pi}}\int_\circ^{\frac{x}{s}}\exp\{-t\}^۲dt+\frac{x}{s}\times\frac{۲}{\sqrt{\pi}}\exp\{-(\frac{x}{s})\}^۲$$

از رابطه فوق یک بار دیگر مشتق گرفته تا مشتق دوم نسبت به x به صورت زیر حاصل شود

$$\begin{aligned}\frac{d^۲}{dx^۲}P(x,s) &= \frac{۱}{s}\times\frac{۲}{\sqrt{\pi}}\exp\{-(\frac{x}{s})\}^۲ + \frac{۱}{s}\times\frac{۲}{\sqrt{\pi}}\exp\{-(\frac{x}{s})\}^۲ \\ &\quad -x\times\frac{۲}{s}\times\frac{۱}{\sqrt{\pi}}\times ۲(\frac{x}{s})\times\frac{۱}{s}\exp\{-(\frac{x}{s})\}^۲ \\ &= \frac{۱}{s}\times\frac{۲}{\sqrt{\pi}}\exp\{-(\frac{x}{s})\}^۲ + \frac{۱}{s}\times\frac{۲}{\sqrt{\pi}}\exp\{-(\frac{x}{s})\}^۲ \\ &\quad -\frac{۴}{x}\times(\frac{x}{s})^۳\times\frac{۱}{\sqrt{\pi}}\exp\{-(\frac{x}{s})\}^۲\end{aligned}$$

۳. با توجه به تعریف آ.۱ در پیوست آ اگر مشتق دوم مثبت نباشد تابع محدب نیست. با توجه به قسمت (۲) قضیه داریم

$$\frac{d^۲}{dx^۲}P(x,s) = \frac{۴}{s\sqrt{\pi}}\exp\left\{-\left(\frac{x}{s}\right)\right\}^۲\left[۱-\frac{x^۲}{s^۲}\right]$$

برای این که $\frac{d^۲}{dx^۲}p(x,s) > ۰$ باید $۱ - \frac{x^۲}{s^۲} > ۰$ که نتیجه می‌شود $x^۲ < s^۲$ و باید به ازای همه مقادیر x و s داشته باشیم $|x| < s$. در حالی که $x \in \mathbb{R}$ است و لذا مشتق دوم همواره مثبت نیست.

۴. با استفاده از آبرامویچ و استاگن[۴] (۲۰۱۲) می‌توان کران‌های زیر را برای تابع خطا نوشت

$$\frac{۱}{x+\sqrt{x^۲+۲}} < \exp\{x\}^۲\int_x^\infty \exp\{-t\}^۲dt \le \frac{۱}{x+\sqrt{x^۲+\frac{۴}{\pi}}}$$

از طرفی تابع خطا به صورت زیر است

$$\begin{aligned}\frac{۱}{x+\sqrt{x^۲+۲}} &< \exp\{\frac{x}{s}\}^۲\int_{\frac{x}{s}}^\infty \exp\{-t\}^۲dt \le \frac{۱}{x+\sqrt{x^۲+\frac{۴}{\pi}}} \\ \frac{۲}{\sqrt{\pi}}\frac{e^{-(\frac{x}{s})^۲}}{(\frac{x}{s})+\sqrt{(\frac{x}{s})^۲+۲}} &< \text{erf}^c\left(\frac{x}{s}\right) \le \frac{۲}{\sqrt{\pi}}\frac{e^{-(\frac{x}{s})^۲}}{(\frac{x}{s})+\sqrt{(\frac{x}{s})^۲+\frac{۴}{\pi}}}\end{aligned}$$

[۴] Abramowitz and Stegun

با استفاده از کران بالای رابطه فوق به ازای x های مثبت داریم

$$\left|x - x\left(۱ - \text{erf}^c\left(\frac{x}{s}\right)\right)\right| = \left|x\text{erf}^c\left(\frac{x}{s}\right)\right|$$

$$\begin{aligned}
\left|x\text{erf}^c\left(\frac{x}{s}\right)\right| &\leq \frac{۲x}{\sqrt{\pi}} \frac{e^{-(\frac{x}{s})^۲}}{\frac{x}{s} + \sqrt{(\frac{x}{s})^۲ + \frac{۴}{\pi}}} = \frac{۲x}{\sqrt{\pi}} \frac{e^{-(\frac{x}{s})^۲}}{\frac{x}{s} + \sqrt{(\frac{x}{s})^۲(۱ + \frac{۴}{\pi}(\frac{s}{x})^۲)}} \\
&\leq \frac{۲}{\sqrt{\pi}} \frac{e^{-(\frac{x}{s})^۲}}{(\frac{x}{s}) + \frac{x}{s}\sqrt{۱ + \frac{۴}{\pi}\frac{s^۲}{x^۲}}} = \frac{۲s}{\sqrt{\pi}} e^{-(\frac{x}{s})^۲} \frac{۱}{۱ + \sqrt{۱ + \frac{۴s^۲}{\pi x^۲}}} \\
&\leq \frac{۲s}{\sqrt{\pi}} e^{-(\frac{x}{s})^۲} \\
&= ۲s\phi(\frac{x}{s}, ۰, \frac{۱}{\sqrt{۲}})
\end{aligned}$$

برای $x < ۰$ روش مشابه‌ای استفاده می‌شود که همان نتیجه را می‌دهد. در نتیجه،

$$||x| - P(x,s)| \leq ۲s\phi\left(\frac{x}{s}, ۰, \frac{۱}{\sqrt{۲}}\right).$$

۵. فرض کنید

$$I = \int_۰^{\frac{x}{s}} \exp(-t^۲)dt.$$

با استفاده از انتگرال‌گیری جز به جز، فرض می‌کنیم

$$\begin{cases} u = e^{-t^۲} \\ dv = dt \end{cases} \Rightarrow \begin{cases} du = -۲te^{-t^۲}dt \\ v = t \end{cases}$$

با توجه به این که

$$\int e^{-t^۲}dt = uv - \int vdu$$

می‌توان نوشت

$$I = te^{-t^۲}|_۰^{\frac{x}{s}} + ۲\int_۰^{\frac{x}{s}} t^۲e^{-t^۲}dt$$

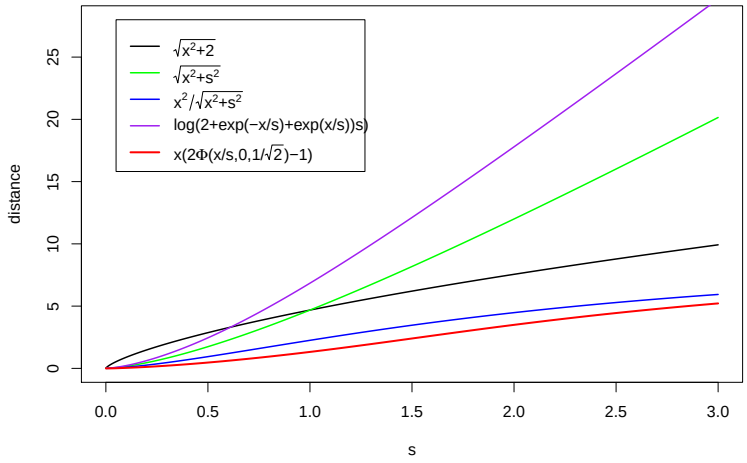
از آن‌جایی که x کوچک است انتگرال صفر شده و نتیجه حاصل می‌شود

$$I = \frac{x}{s}\exp\{-(\frac{x}{s})^۲\} + ۲\int_۰^{\frac{x}{s}} t^۲e^{-t^۲}dt$$

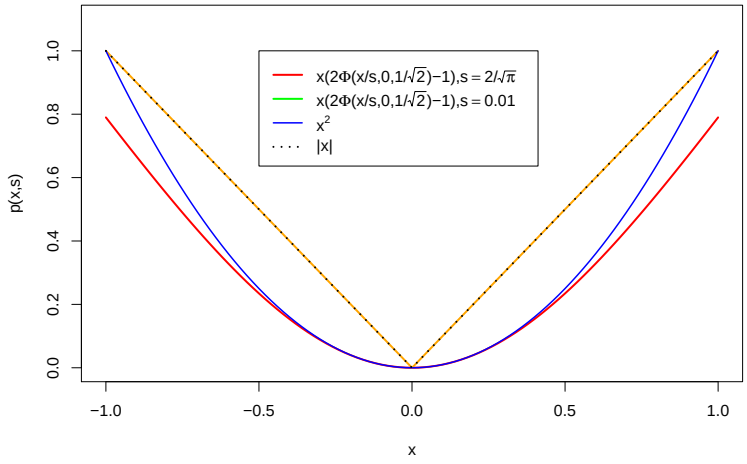
□

سمت راست نامساوی (۴.۲) وقتی $s \to \infty$، با سرعت به نمایی صفر نزدیک می‌شود. شکل ۱.۲ (الف) این نتیجه را به خوبی نشان می‌دهد. همان طور که از شکل مشخص است، تابع پیشنهادی به تابع قدر مطلق سریع‌تر از سایر توابع همگرا می‌شود. نمودار (ب) در شکل ۱.۲ نشان می‌دهد که وقتی s مقدار کوچکی است، تابع پیشنهادی مطابق قدرمطلق رفتار می‌کند و به ازای $s = \frac{۲}{\sqrt{\pi}} \approx ۱$، تابع مانند تابع $x^۲$ در مجاورت $x = ۰$ رفتار می‌کند.

(الف) اختلاف بین $p(x, s)$ و $|x|$



(ب) دو مثال از تابع جریمه DLASSO



شکل ۱.۲: ویژگی‌های کلیدی جریمه DLASSO. (الف) همگرایی سریع به تابع قدرمطلق در مقایسه با تقریب‌های دیگر. (ب) تابع جریمه DLASSO برای sهای کوچک شبیه لاسو و به ازای $s = \frac{۲}{\sqrt{\pi}}$ شبیه ریج است.

نکته نهایی بحث در مورد پیچیدگی محاسباتی است، که تنها مشکل بالقوه پیشنهاد ما است. با این حال، تعدادی از تقریب‌های خوب و سریع که در نوشته‌ها برای ارزیابی ارائه شده‌اند از تابع خطا و توزیع تابع چگالی نرمال هستند. یک نکته خوب استفاده از تقریب تیلور (پیوست آ.۱.۱)است. به عنوان مثال، تقریب‌ها می‌توانند بر اساس یکی از عبارت‌های زیر باشند

$$\text{erf(x)} = \frac{۲x}{\sqrt{\pi}} \sum_{j=۰}^{\infty} \frac{(-۱)^j x^{۲j}}{j!(۲j+۱)} \tag{۵.۲}$$

$$\text{erf(x)} = \frac{۲xe^{-x^۲}}{\sqrt{\pi}} \sum_{j=۰}^{\infty} \frac{۲^j x^{۲j}}{۱\cdot۳\cdot۵\cdots(۲j+۱)}, \tag{۶.۲}$$

$$\text{erf}^c(x) \approx \frac{e^{-x^۲}}{x\sqrt{\pi}} \sum_{j=۰}^{k} (-۱)^j \frac{(۲j)!}{j!} (۲x)^{-۲j}. \tag{۷.۲}$$

برای $|x|$ کوچک، سری (۵.۲) نسبت به سری (۶.۲) سریع‌تر است، زیرا نیازی نیست به صورت نمایی محاسبه شود. اگرچه، سری (۶.۲) برتر از سری (۵.۲) برای $|x|$ است. زیرا شامل هیچ حذفی نمی‌شود. برای $|x|$ بزرگ، هیچ یک از سری‌ها رضایت‌بخش نیستند و در این مورد بهتر است از سری برتر (۷.۲) در انبساط مجانبی برای متمم تابع خطا استفاده شود.

تقریب‌های بسط تیلور، الگوریتم‌های تناوبی سریعی برای تقریب تابع خطا یا تابع چگالی نرمال دارد. به عنوان مثال، بورجسون و ساندبرگ [۵] (۱۹۷۹)، چویلارد و رول[۶] (۲۰۰۸)، کودی[۷] (۱۹۹۰)، لی[۸] (۱۹۹۲)، اولور[۹] (۲۰۱۰)، پرس[۱۰] (۲۰۱۰) و لئال [۱۱] و همکاران (۲۰۱۲) را نگاه کنید. به طور خاص دو تقابل سریع وجود دارند که به صورت زیر هستند:

$$\text{erf}^c(x) \approx \tanh\left(\frac{۳۹x}{۲\sqrt{\pi}} - \frac{۱۱۱}{۲}\arctan\left(\frac{۳۵x}{۱۱۱\sqrt{\pi}}\right)\right),$$

$$\Phi(x,۰,۱) \approx \left(\frac{۱}{۱/۹\sqrt{\pi}}\left(\sin\left(\frac{\pi x}{۱۰}\right)+\sin(x)\right)+۰/۵\right) I_{(|x|\le ۱/۵۱۳۸۵۹)}$$
$$+\left(۱-e^{-۱/۷۸}+\frac{x}{e^{x+۱۰}}\right) I_{(x>۱/۵۱۳۸۵۹)} + \left(۱-e^{-۱/۷۸|x|}-\frac{|x|}{e^{|x|+۱۰}}\right) I_{(x<۱/۵۱۳۸۵۹)}$$

[۵] Borjesson and Sundberg
[۶] Chevillard and Revol
[۷] Cody
[۸] Lee
[۹] Olver
[۱۰] Press
[۱۱] Leal .

اینها نه تنها سریع بلکه بسیار دقیق هستند، ماکزیمم خطای مطلق آخرین تابع $۱۰^{-۴}$ است. برای آگاهی بیشتر حاصلی‌مشهدی و وینچوتی (۲۰۱۶) را ببینید.

۳.۲ برآوردگرلاسوی مشتق‌پذیر

با جای‌گذاری تابع جریمه لاسوی مشتق‌پذیر (۱.۲) در معادله (۱۱.۱)، برآوردگر لاسوی مشتق‌پذیر به صورت زیر بدست می‌آید

$$\hat{\boldsymbol{\beta}}^{DLASSO} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\operatorname{argmin}}\left\{ S(\boldsymbol{\beta}) + \lambda \sum_{j=۱}^{p} P(\beta_j, s) \right\}. \qquad (۸.۲)$$

برای درک بهتری از برآوردگر نتیجه‌شده، مدل زیر را در نظر بگیرید

$$y_i = \beta_i + \epsilon_i \qquad i = ۱, \ldots, n$$

که در آن $\epsilon_i \sim N(۰, \sigma^۲)$. در این صورت $\hat{\beta}_i^{DLASSO}$ پاسخ دستگاه معادله‌های زیر است که نسبت به β_i مشتق می‌گیریم

$$L(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{i=۱}^{p} \beta_i \left(۲\Phi(\frac{\beta_i}{s}, ۰, \frac{۱}{\sqrt{۲}}) - ۱ \right) \qquad s > ۰, \lambda \geq ۰$$

$$\frac{d}{d\beta_i} L(\boldsymbol{\beta}) = \lambda \left(۲\Phi(\frac{\beta_i}{s}, ۰, \frac{۱}{\sqrt{۲}}) - ۱ + ۲(\frac{\beta_i}{s})\phi(\frac{\beta_i}{s}, ۰, \frac{۱}{\sqrt{۲}}) \right) - ۲(y_i - \beta_i) = ۰, i = ۱, \ldots, n$$

شکل ۲.۲ برآوردگرها را به ازای مقادیر مختلف s نشان می‌دهد. همان طور که قبلا اشاره شد و با توجه به این شکل، (الف)، (ج) و (د) به ترتیب توابع مشابه لاسو، ریج و جریمه‌نشده را نشان می‌دهد.

با توجه به اینکه برای مقادیر کوچک s $(s \to ۰)$، تقریب‌های بسیار خوبی برای تابع قدر مطلق به دست می‌آید، انتظار داریم که در این حالت برآوردگرها، ویژگی‌هایی شبیه برآوردگر لاسو داشته باشند که در قضیه‌های بعدی اثبات شده است.

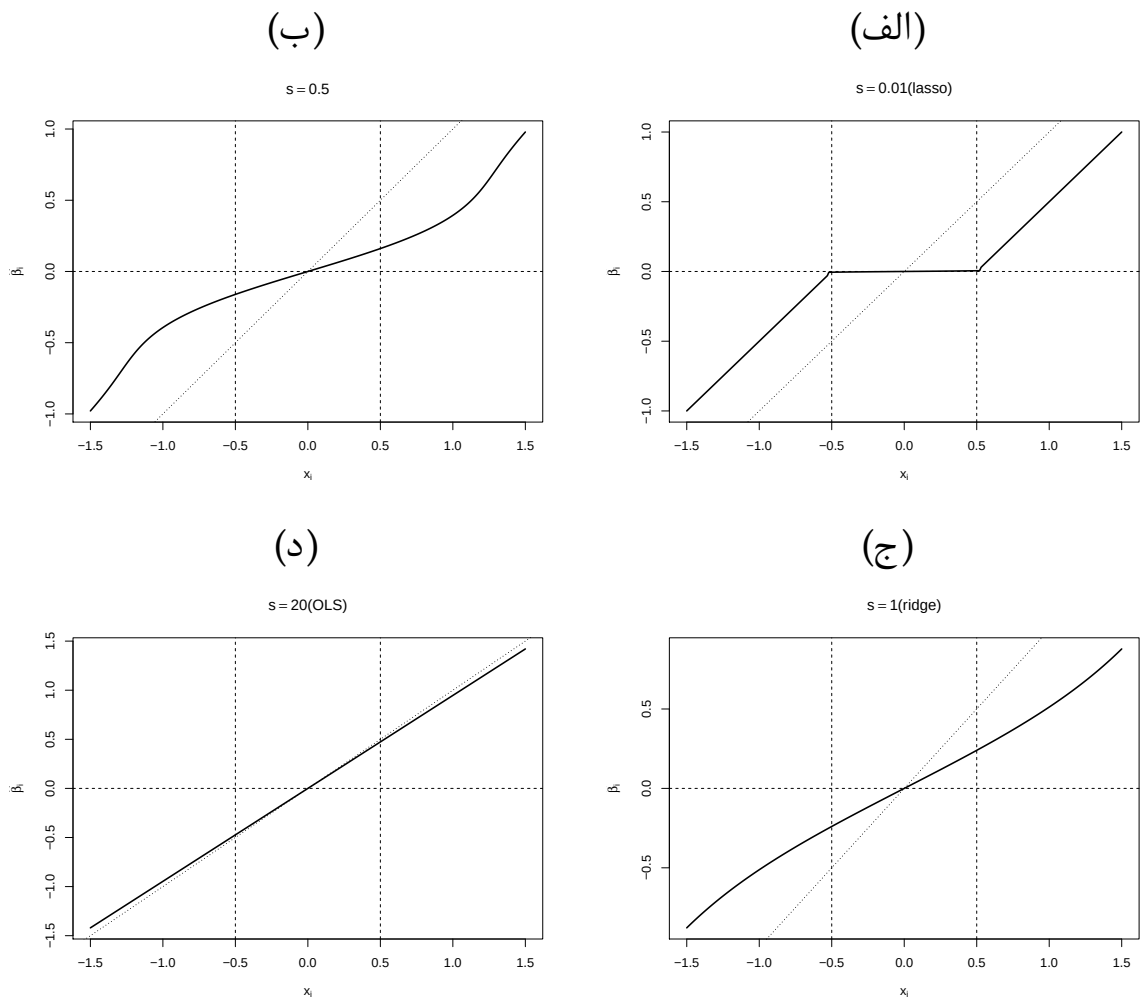
قضیه ۱.۳.۲. برای هر $u \in \mathbb{R}$، $\lambda \geq ۰$ و $s > ۰$ ، تعریف کنید

$$k(\boldsymbol{u}, s) = L(\boldsymbol{\beta} + \boldsymbol{u}) - L(\boldsymbol{\beta}) \qquad (۹.۲)$$

که در آن $L(\boldsymbol{\beta})$ تابع هدف در معادله (۸.۲) است. در این صورت

$$\lim_{s\to ۰} k(u, s) = \boldsymbol{u}^\top \boldsymbol{X}_n^\top \boldsymbol{X}_n \boldsymbol{u} - ۲\boldsymbol{u}^\top N\left(۰, \sigma^۲(\boldsymbol{X}_n^\top \boldsymbol{X}_n) \right) + \lambda \sum_{i=۱}^{m} \left(|u_i| I(\beta_i = ۰) + u_i \text{sign}(\beta_i + u_i) \right),$$

که در آن $N(۰, \sigma^۲)$ یک متغیر تصادفی با توزیع نرمال است.



شکل ۲.۲: نمودار توابع آستانه‌ای به‌ازای $\lambda = ۱$ و مقادیر مختلف s: $s = ۰/۰۱$ (تابع لاسو)، $s = ۰/۵$، $s = ۱$ (تابع ریج)، $s = ۲۰$ (برآوردگر OLS)

برهان. باجای گذاری در $L(\boldsymbol{\beta})$ داریم،

$$
\begin{aligned}
k(\boldsymbol{u}, s) &= L(\boldsymbol{\beta} + \boldsymbol{u}) - L(\boldsymbol{\beta}) \\
&= (\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta} + \boldsymbol{u}))^\top (\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta} + \boldsymbol{u})) + \lambda \sum_{i=۱}^{p} (\beta_i + u_i) \left(۲\Phi\big(\frac{(\beta_i + u_i)}{s}, ۰, \frac{۱}{\sqrt{۲}}\big) \right) \\
&\quad - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{i=۱}^{p} \beta_i \left(۲\Phi\big(\frac{\beta_i}{s}, ۰, \frac{۱}{\sqrt{۲}}\big) - ۱ \right) \\
&= (\epsilon - \boldsymbol{X}\boldsymbol{u})^\top (\epsilon - \boldsymbol{X}\boldsymbol{u}) - \epsilon^\top \epsilon + ۲\lambda \sum_{i=۱}^{p} \beta_i \int_{\frac{\beta_i}{s}}^{\frac{\beta_i + u_i}{s}} \frac{۱}{\sqrt{\pi}} e^{(-t)^۲} dt \\
&\quad + \lambda \sum_{i=۱}^{p} u_i \left(۲\Phi\big(\frac{\beta_i + u_i}{s}, ۰, \frac{۱}{\sqrt{۲}}\big) - ۱ \right)
\end{aligned}
$$

حال اگر $s \to ۰$ داریم

$$\begin{aligned}
\lim_{s\to ۰} k(\boldsymbol{u}, s) &= \lim_{s\to ۰} L(\boldsymbol{\beta}+\boldsymbol{u}) - \lim_{s\to ۰} L(\boldsymbol{\beta}) \\
&= (\epsilon - \boldsymbol{X}\boldsymbol{u})^\top(\epsilon - \boldsymbol{X}\boldsymbol{u}) - \epsilon^\top\epsilon + \lim_{s\to ۰}\Big\{ ۲\lambda \sum_{i=۱}^{p} \beta_i \int_{\frac{\beta_i}{s}}^{\frac{\beta_i+u_i}{s}} \frac{۱}{\sqrt{\pi}} e^{(-t)^۲} dt \\
&\quad + \lambda \sum_{i=۱}^{p} u_i \big(۲\Phi(\frac{\beta_i+u_i}{s}, ۰, \frac{۱}{\sqrt{۲}}) - ۱ \big) \Big\} \\
&= \boldsymbol{u}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{u} - ۲\boldsymbol{u}^\top N\big(۰, \sigma^۲(\boldsymbol{X}^\top\boldsymbol{X})\big) + \lim_{s\to ۰} \lambda \sum_{i=۱}^{p} \Big(۲\beta_i \frac{u_i}{s} \phi(\frac{u_i}{s}, ۰, \frac{۱}{\sqrt{۲}}) \\
&\quad + \lim_{s\to ۰}\Big\{ u_i \Big(۲\Phi(\frac{\beta_i+u_i}{s}, ۰, \frac{۱}{\sqrt{۲}}) - ۱ \Big) \Big\} \Big) \\
&= \boldsymbol{u}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{u} - ۲\boldsymbol{u}^\top N\big(۰, \sigma^۲(\boldsymbol{X}^\top\boldsymbol{X})\big) \\
&\quad + \lim_{s\to ۰} \lambda \sum_{i=۱}^{p} \begin{cases} u_i \Big(۲\Phi\big(\frac{u_i}{s}, ۰, \frac{۲}{\sqrt{۲}}\big) - ۱ \Big) & \beta_i = ۰ \\ ۲\beta_i \frac{u_i}{s} \phi\big(\frac{u_i}{s}, ۰, \frac{۱}{\sqrt{۲}}\big) + \boldsymbol{u}_i \Big(۲\Phi\big(\frac{\beta_i+u_i}{s}, ۰, \frac{۱}{\sqrt{۲}}\big) - ۱ \Big) & \beta_i \neq ۰ \end{cases} \\
&= \boldsymbol{u}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{u} - ۲\boldsymbol{u}^\top N\big(۰, \sigma^۲(\boldsymbol{X}^\top\boldsymbol{X})\big) \\
&\quad + \lambda \sum_{i=۱}^{p} \begin{cases} |u_i| & \beta_i = ۰ \\ u_i & \beta_i + u_i > ۰, \\ -u_i & \beta_i + u_i < ۰ \end{cases}
\end{aligned}$$

و داریم

$$k(u) = \boldsymbol{u}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{u} - ۲\boldsymbol{u}^\top N\big(۰, \sigma^۲(\boldsymbol{X}^\top\boldsymbol{X})\big) + \lambda \sum_{i=۱}^{p} \Big(|u_i| I(\beta_i = ۰) + u_i \text{sign}(\beta_i + u_i) \Big).$$

□

در قضیه بعد نشان می‌دهیم که حداقل سرعت s که همگرایی برآوردگرها را به لاسو تضمین می‌کند، $n^{(-\frac{۱}{۲}+\epsilon)}$ به ازای هر $\epsilon > ۰$ است.

قضیه ۲.۳.۲. فرض کنید $\boldsymbol{\beta}$ مجموعه‌ای از ضرایب تنک باشد، $\boldsymbol{u} \in \mathbb{R}^p$، $s_n = s/(n^{\frac{۱}{۲}+\epsilon}) \to ۰$، $\epsilon > ۰$، $\lambda_n/\sqrt{n} \to \lambda_۰ \geq ۰$، و $\boldsymbol{X}^\top\boldsymbol{X}/n \to \boldsymbol{\Sigma}$ که $\boldsymbol{\Sigma}$ معکوس‌پذیر است. در این صورت، $\sqrt{n}(\hat{\boldsymbol{\beta}}^{DLASSO} - \boldsymbol{\beta}) \to \underset{\boldsymbol{u}}{\text{argmin}} k(\boldsymbol{u})$ که در آن

$$k(\boldsymbol{u}) = -۲\boldsymbol{u}^\top N\big(۰, \sigma^۲\boldsymbol{\Sigma}\big) + \boldsymbol{u}^\top \Sigma \boldsymbol{u} + \lambda_۰ \sum_{i=۱}^{p} \{ u_i \text{sign}(\beta_i) I(\beta_i \neq ۰) + |u_i| I(\beta_i = ۰) \}.$$

برهان. فرض کنید $k_n(\boldsymbol{u}) = L(\boldsymbol{\beta} + \frac{\boldsymbol{u}}{\sqrt{n}}) - L(\boldsymbol{\beta})$. در این صورت

$$
\begin{aligned}
k_n(\boldsymbol{u}) &= L(\boldsymbol{\beta} + \frac{\boldsymbol{u}}{\sqrt{n}}) - L(\boldsymbol{\beta}) \\
&= (\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta} + \frac{\boldsymbol{u}}{\sqrt{n}}))^\top (\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta} + \frac{\boldsymbol{u}}{\sqrt{n}})) \\
&\quad + \lambda_n \sum_{i=1}^{p} (\beta_i + \frac{u_i}{\sqrt{n}}) \Big(2\Phi(\frac{(\beta_i + \frac{u_i}{\sqrt{n}})}{s}, 0, \frac{1}{\sqrt{2}}) \Big) \\
&\quad - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_n \sum_{i=1}^{p} \beta_i \Big(2\Phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) \\
&= (\epsilon - \boldsymbol{X}\frac{\boldsymbol{u}}{\sqrt{n}})^\top (\epsilon - \boldsymbol{X}\frac{\boldsymbol{u}}{\sqrt{n}}) - \epsilon^\top \epsilon + 2\lambda_n \sum_{i=1}^{p} \beta_i \int_{\frac{\beta_i}{s}}^{\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\
&\quad + \lambda_n \sum_{i=1}^{p} \frac{u_i}{\sqrt{n}} \Big(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) \\
&= \boldsymbol{u}^\top \frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} \boldsymbol{u} - 2\boldsymbol{u}^\top N(0, \sigma^2(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n})) + 2\lambda_n \sum_{i=1}^{p} \beta_i \int_{\frac{\beta_i}{s}}^{\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\
&\quad + \lambda_n \sum_{i=1}^{p} \frac{u_i}{\sqrt{n}} \Big(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big)
\end{aligned}
$$

با توجه به فرضیات اگر $n \to \infty$ ، داریم

$$
\begin{aligned}
k_n(\boldsymbol{u}) &\xrightarrow{n\to\infty} \boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u} - 2\boldsymbol{u}^\top N(0, \sigma^2 \boldsymbol{\Sigma}) + \lim_{n\to\infty} 2\lambda_n \sum_{i=1}^{p} \beta_i \int_{\frac{\beta_i}{s}}^{\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\
&\quad + \lim_{n\to\infty} \lambda_n \sum_{i=1}^{p} \frac{u_i}{\sqrt{n}} \Big(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) \\
&= \boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u} - 2\boldsymbol{u}^\top N(0, \sigma^2 \boldsymbol{\Sigma}) + 2\lambda_0 \sum_{i=1}^{p} \beta_i \frac{u_i}{s\sqrt{\pi}} \lim_{n\to\infty} e^{-(\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s})^2} \\
&\quad + \lambda_0 \sum_{i=1}^{p} u_i \Big(\lim_{n\to\infty} 2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) \\
&= \boldsymbol{u}^\top \Sigma \boldsymbol{u} - 2\boldsymbol{u}^\top N(0, \sigma^2 \Sigma) \\
&\quad + \lim_{n\to\infty} \begin{cases} \lambda_0 \sum_{i=1}^{p} u_i \Big(2\Phi(\frac{\frac{u_i}{\sqrt{n}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) & \beta_i \in S_0 \\ 2\lambda_0 \sum_{i=1}^{p} \Big(\beta_i \frac{u_i}{s\sqrt{\pi}} e^{-(\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s})^2} + u_i \Big(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{n}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) \Big) & \beta_i \in S_0^c \end{cases}
\end{aligned}
$$

که S_0 و S_0^c به ترتیب مجموعه ضرایب صفر و غیر صفر هستند. □

مشابه روش نایت و فو [۱۲] (۲۰۰۰)، می‌توان نشان داد که این نتایج برآورد تنک پارامترها را تضمین می‌کند. در اثبات قضیه ۲.۳.۲ فرض شده است که $S_n\sqrt{n} \to 0$. بنابراین، اگر مقدار S_n را کمتر از $\frac{1}{\sqrt{n}}$ انتخاب شود نتایج بسیار شبیه لاسو خواهد بود.

۴.۲ الگوریتم برای برآورد پارامترها

درستنمایی جریمه (۸.۲) نسبت به $\boldsymbol{\beta}$ مشتق‌پذیر است. بنابراین روش‌های بهینه‌سازی استاندارد می‌تواند برای یافتن مینیمم آن استفاده شود. با این وجود، سرعت یافتن پاسخ در این روش‌ها به کندی خواهد بود.
در این بخش الگوریتم کارایی بر اساس ویژگی مشتق‌پذیری تابع جریمه لاسوی مشتق پذیر ارائه می‌کنیم. بدین منظور، مشابه روش پیشنهادی فن و لی (۲۰۰۱) الگوریتم تکرارشونده زیر پیشنهاد می‌شود

$$\boldsymbol{\beta}^{(k)} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Sigma}(\boldsymbol{\beta}^{(k-1)}, \lambda, s)\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \qquad k = 1, 2, \ldots \qquad (10.2)$$

در اینجا $\boldsymbol{\beta}^{(0)}$ یک برآورد اولیه برای پارامترهاست که می‌تواند برآورد لاسو باشد و $\boldsymbol{\Sigma}(\boldsymbol{\beta}^{(k-1)}, \lambda, s)$ به صورت زیر تعریف می شود

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\beta}^{(k-1)}, \lambda, s) = \lambda Diag\Bigg[&\left(2\Phi\left(\frac{\beta_i^{(k-1)}}{s}, 0, \frac{1}{\sqrt{2}}\right) - 1\right) \\ &+ \frac{\beta_i^{(k-1)}}{s}\phi\left(\frac{\beta_i^{(k-1)}}{s}, 0, \frac{1}{\sqrt{2}}\right) / \beta_i^{(k-1)}, i = 1, \ldots, r\Bigg].\end{aligned}$$

به منظور به دست آوردن این رابطه، از تقریب تیلور مرتبه اول برای تابع جریمه لاسوی مشتق‌پذیر حول $\boldsymbol{\beta}^{(0)}$ استفاده می‌کنیم. به عبارت دیگر،

$$\begin{aligned}\boldsymbol{\beta}\left(2\Phi(\frac{\boldsymbol{\beta}}{s}, 0, \frac{1}{\sqrt{2}}) - 1\right) \approx \boldsymbol{\beta}^{(0)}&\left(2\Phi(\frac{\boldsymbol{\beta}^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1\right) \\ &+ \Bigg(2\Phi(\frac{\boldsymbol{\beta}^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \\ &+ \frac{2\boldsymbol{\beta}^{(0)}}{s}\phi(\frac{\boldsymbol{\beta}^{(0)}}{s}, 0, \frac{1}{\sqrt{2}})\Bigg)(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}).\end{aligned}$$

توجه داشته باشید که مشتق‌پذیر بودن تابع لاسو بدین معناست که نیازی به تقریب درجه دوم موضعی وجود ندارد (فن و لی، ۲۰۰۱). اگر $\boldsymbol{\beta} \approx \boldsymbol{\beta}^{(0)}$، می‌توانیم معادله را

[۱۲] Knight and Fu

به صورت زیر بازنویسی کنیم

$$\begin{aligned}\boldsymbol{\beta}\Big(۲\Phi(\frac{\boldsymbol{\beta}}{s},۰,\frac{۱}{\sqrt{۲}})-۱\Big) \approx \boldsymbol{\beta}^{(۰)}\Big(۲\Phi(\frac{\boldsymbol{\beta}^{(۰)}}{s},۰,\frac{۱}{\sqrt{۲}})-۱\Big)\\ +\frac{۱}{\boldsymbol{\beta}^{(۰)}}\Big(۲\Phi(\frac{\boldsymbol{\beta}^{(۰)}}{s},۰,\frac{۱}{\sqrt{۲}})-۱\\ +\frac{۲\boldsymbol{\beta}^{(۰)}}{s}\phi(\frac{\boldsymbol{\beta}^{(۰)}}{s},۰,\frac{۱}{\sqrt{۲}})\Big)(\boldsymbol{\beta}^{۲}-\boldsymbol{\beta}^{(۰)^{۲}}). \quad (۱۱.۲)\end{aligned}$$

با جایگذاری(۱۱.۲) در تابع جریمه (۸.۲) داریم

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\top}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \;+\; \lambda\sum_{i=۱}^{p}\Big[\beta_i^{(۰)}\Big(۲\Phi(\frac{\beta_i^{(۰)}}{s},۰,\frac{۱}{\sqrt{۲}})-۱\Big)\\ +\;\frac{۱}{\beta_i^{(۰)}}\Big(۲\Phi(\frac{\beta_i^{(۰)}}{s},۰,\frac{۱}{\sqrt{۲}})-۱\\ +\;\frac{۲\beta_i^{(۰)}}{s}\phi(\frac{\beta_i^{(۰)}}{s},۰,\frac{۱}{\sqrt{۲}})\Big)(\beta_i^{۲}-\beta_i^{(۰)^{۲}})\Big] \quad (۱۲.۲)\end{aligned}$$

در این صورت، مقدار بهینه با روش محاسبات تکراری رگرسیون ریج (۱۰.۲) به دست می‌آید. این الگوریتم در بسته $DLASSO$ در نرم‌افزار R قابل اجرا است. این بسته اجازه می‌دهد پارامترهای تنظیم‌کننده[۱۳] λ و s را با معیاری که انتخاب مدل رایج مانند AIC، BIC یا GCV انتخاب کنیم.

۵.۲ مطالعه شبیه‌سازی

در این بخش، برای بررسی کارآیی تابع جریمه در رگرسیون، از شبیه‌سازی استفاده شده است. مدل زیر را در نظر بگیرید

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\sigma\epsilon \qquad \epsilon\sim N(۰,۱)$$

مشابه ژو و هستی (۲۰۰۵)، سه سناریوی مختلف زیر را طراحی می‌کنیم
سناریوی ۱ (استاندارد): بردار پارامتر $\boldsymbol{\beta}$ را به صورت $(۳,۱.۵,۰,۰,۲,۰,۰,۰)$ قرار می‌دهیم. متغیرهای $\boldsymbol{X}$ را از توزیع نرمال چند متغیره با میانگین صفر و ماتریس همبستگی $Cor(\boldsymbol{X}_i\boldsymbol{X}_j)=۰/۵^{|i-j|}$ و براین اساس به ازای $\boldsymbol{\sigma}^{۲}=۳$، متغیر پاسخ را تولید می‌کنیم.
سناریوی ۲ ($\boldsymbol{\beta}$های کوچک): مشابه سناریوی ۱ با این تفاوت که بردار $\boldsymbol{\beta}$ را به صورت $\beta_j=۰/۵$ به ازای $j=۱,۲,\ldots,۸$ در نظر می‌گیریم.

[۱۳] Tuning Parameters

سناریوی ۳ (متغیرهای پیشگو همبسته): در این سناریو فرض می‌کنیم $p = ۳$ و ضرایب را به سه گروه تقسیم می‌کنیم. $\boldsymbol{\beta}^{(۱)} = (۱, ۲, ۳, ۴, ۵)$، $\boldsymbol{\beta}^{(۲)} = (۰/۵, ۰/۵, ۰/۵, ۰/۵, ۰/۵)$ و $\boldsymbol{\beta}^{(۳)} = (۰, ۰, ۰, ۰, ۰)$. همبستگی بالای ۰/۹ را بین هر جفت از اولین پنج متغیر در نظر می‌گیریم. به طور مشابه، همبستگی ۰/۵ را در گروه دوم و هیچ وابستگی را در گروه سوم در نظر می‌گیریم. در نهایت $\boldsymbol{\sigma}^{۲} = ۱۵$.

سناریو ۴ (استاندارد) : بردار پارامتر $\boldsymbol{\beta}$ را به صورت $(۳, ۱.۵, ۰, ۰, ۲, \underbrace{۰, ۰, ۰, \cdots, ۰}_{۲۹۵})$ قرار می‌دهیم. متغیرهای $\boldsymbol{X}$ را از توزیع نرمال چند متغیره با میانگین صفر و ماتریس همبستگی $Cor(\boldsymbol{X}_i \boldsymbol{X}_j) = ۰/۵^{|i-j|}$ و براین اساس به ازای $\boldsymbol{\sigma}^{۲} = ۳$، متغیر پاسخ را تولید می‌کنیم.

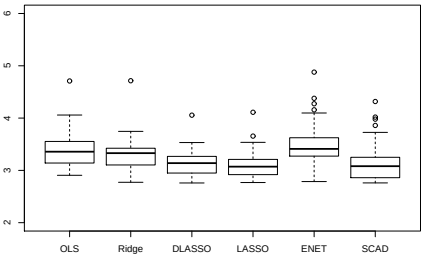
برای هر سه سناریو، ۵۰ مجموعه داده از ۲۴۰ مشاهده تولید می‌کنیم. ۴۰ مشاهده از آن‌ها را به عنوان مجموعه آموزشی و b مشاهده را به عنوان مجموعه آزمون در نظر می‌گیریم. پارامترهای جریمه بر اساس روش اعتبارسنجی متقابل ۱۰ تایی روی MSE محاسبه می‌شوند.

مدل‌های زیر را مقایسه می‌کنیم: لاسوی مشتق‌پذیر (به ازای $s = ۰/۰۱$)، لاسوی مشتق‌پذیر (با s و λ محاسبه‌شده براساس CV)، لاسو، ریج، الاستیک‌نت (enet)، اسکد و کمترین توان‌های دوم (LS).

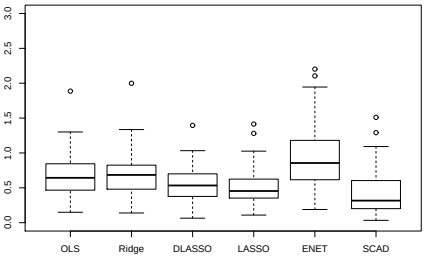
برای محاسبه برآوردگرهای لاسو و الاستیک‌نت از بسته msgps (هیروس، ۲۰۱۲) استفاده می‌کنیم. از بسته ncvreg (برهنی و هوانگ، ۲۰۱۱) و از بسته glmnet (فریدمن و همکاران، ۲۰۱۰) به ترتیب برای یافتن برآوردگرهای اسکد و ریج استفاده شده است. هم‌چنین برای روش‌های dlasso از بسته DLASSO (حاصلی مشهدی، ۲۰۱۷) استفاده کرده‌ایم.

شکل ۳.۲ نتایج را نشان می‌دهد. برای هر سناریو، توزیع (روی ۵۰ تکرار) از MSE روی مجموعه آزمون که به صورت $\sum_{i=۱}^{۲۰۰} \frac{(y_i - \hat{y}_i)^۲}{۲۰۰}$ تعریف می‌شود و MSE پارامترهای برآوردشده که به صورت $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top S_X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ تعریف می‌شود که در آن $\boldsymbol{\beta}$ و $\hat{\boldsymbol{\beta}}$ به ترتیب پارامترهای واقعی و مقدار برآوردشده است.

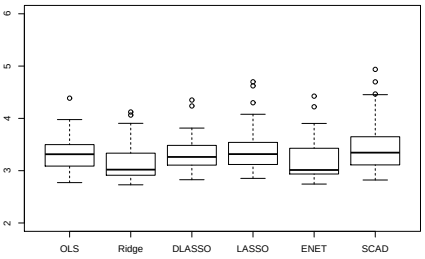
سناریوی ۱ : MSE(Prediction)



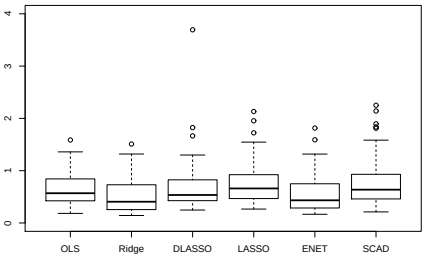
سناریوی ۱: MSE(Beta)



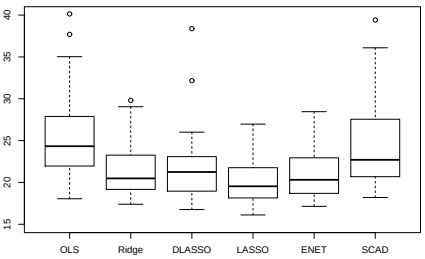
سناریوی ۲ : MSE(Prediction)



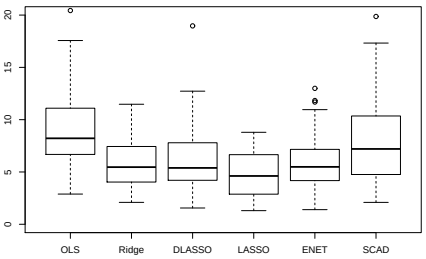
سناریوی ۲: MSE(Beta)



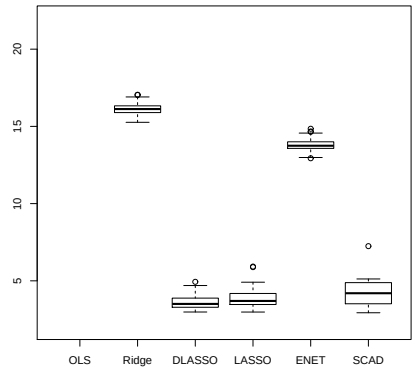
سناریوی ۳ : MSE(Prediction)



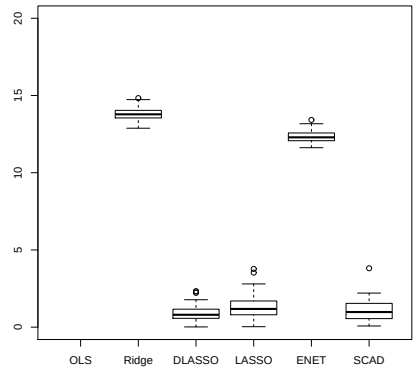
سناریوی ۳: MSE(Beta)



سناریوی ۴ : MSE(prediction)



سناریوی ۴ : MSE(Beta)



شکل ۳.۲: مطالعه شبیه‌سازی مقایسه برآوردگر لاسوی مشتق‌پذیر با روش‌های موجود در چهار سناریوی مختلف

در اغلب سناریوها، جریمه لاسوی مشتق‌پذیر بهتر یا مشابه جریمه‌های موجود و بهتر از جریمه ریج رفتار می‌کند که تنها تابع جریمه‌ی مشتق‌پذیر در بین سایر تابع جریمه‌های رایج است و همچنین نسبت به اسکد که تنها جریمه غیر محدب است (به جز سناریو ۱) برتری دارد. اگرچه جریمه لاسو نیز در بعضی اوقات بهتر از لاسوی مشتق‌پذیر است به یاد داشته باشیم که ایرار تابع جریمه لاسو مشتق‌پذیر نبودن آن در صفر بوده که دلیل اصلی پیشنهاد تابع جریمه لاسوی مشتق‌پذیر است.

فصل ۳

لاسوی مشتق‌پذیر در مدل تفاضلی

۱.۳ مقدمه

مدل‌های نیمه‌پارامتری شامل قسمت پارامتری و قسمت نیمه‌پارامتری هستند. این مدل‌ها کاربردهای فراوانی دارند. انگل[۱] و همکاران (۱۹۸۶) اولین‌ها بودند که مدل خطی نیمه‌پارامتری را به کار بردند. در مدل رگرسیونی نیمه‌پارامتری، یاچینو [۲] (۲۰۰۳) بر برآورد پارامتر خطی متمرکز شد و از روش تفاضل برای از بین بردن اریبی ناشی از حضور پارامتر ناپارامتری در مدل استفاده نمود. روش تفاضلی کردن برای حذف اثر تابع ناپارامتری در مدل‌های رگرسیون ناپارامتری چندان جدید نیست (به رایس[۳](۱۹۸۴) مراجعه شود). در حالی که به کار بردن این ایده در مدل‌های نیمه پارامتری نسبتا جدیدتر hsj (به آن و پاول[۴]، ۱۹۹۳ و یاچیو، ۱۹۹۷، ۱۹۹۸، ۱۹۹۹، ۲۰۰۰ و ۲۰۰۱، تاباکان و آکدنیز[۵]، ۲۰۱۰ ، روزبه و همکاران، ۲۰۱۰، آرشی و همکاران، ۲۰۱۱ و روزبه، ۱۳۹۰ مراجعه شود). در این فصل علاقه‌مندیم برآوردگر لاسوی مشتق‌پذیر را با مدل

[۱] Engle
[۲] Yatchew
[۳] Rice
[۴] Ahn and Pawell
[۵] Tabakan and Akdeniz

رگرسیون نیمه‌پارامتری به‌دست بیاوریم. مطالب این فصل عمدتا برگرفته از پایان‌نامه بغلانی (۱۳۹۴) است اما بخش لاسوی مشتق‌پذیر آن جدید بوده و حاصل تحقیقات نویسنده این مجموعه می‌باشد.

۲.۳ رگرسیون ناپارامتری

در مدل رگرسیون ناپارامتری n جفت از مشاهدات $(t_1, y_1), \ldots, (t_n, y_n)$ در دست داریم. متغیر پاسخ y توسط معادله رگرسیون ناپارامتری که به t وابسته است به صورت زیر خواهد بود

$$y_i = f(t_i) + \epsilon_i, \qquad E(\epsilon_i) = 0, i = 1, \ldots, n \tag{۱.۳}$$

به طوری که $f(\cdot)$ یک تابع ناپارامتری است و t_i ها از هم مستقل هستند. می‌خواهیم تابع $f(.)$ را تحت فرضیات زیر برآورد کنیم:

(۱) t_i ها عضو یک مجموعه فشرده باشند.

(۲) تابع $f(\cdot)$ یک تابع هموار باشد (مشتقات اول و دوم وجود داشته باشند).

برآوردگر $f(t)$ با $\hat{f}_n(t)$ نشان داده می‌شود. همچنین $\hat{f}_n(t)$ به عنوان یک هموارساز شناخته می‌شود. در ادامه برای برآورد $f(t)$ به معرفی هموارساز خطی[۶] و هموارساز میانگین متحرک[۷] و همچنین هموارسازهای کرنل[۸] می‌پردازیم.

۳.۳ هموارساز خطی

یک برآوردگر $\hat{f}_n$ یک هموارساز خطی است اگر برای هر t، یک بردار $\boldsymbol{W}(t) = (w_1(t), \ldots, w_n(t))^\top$ وجود داشته باشد به طوری که

$$\hat{f}_n = \sum_{i=1}^{n} W_i(t) y_i. \tag{۲.۳}$$

یک بردار از مقادیر برازش‌شده به صورت زیر تعریف می‌شود:

$$\boldsymbol{f} = (\hat{f}_1(t_1), \ldots, \hat{f}_n(t_n)). \tag{۳.۳}$$

[۶] Linear Smoother
[۷] Moving Average Smoother
[۸] Kernel Smoothers

به طوری که $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. بنابراین : $\boldsymbol{f} = \boldsymbol{W}\boldsymbol{y}$ به طوری که $\boldsymbol{W}$ یک ماتریس $n \times n$ است که i امین ردیف آن $W(t_i)^\top$ می‌باشد. لذا قرار می‌دهیم

$$W_{ij} = W_j(t_i).$$

اعضای i امین سطر نشان‌دهنده وزن‌هایی است که به هر y_i داده می‌شود.
وزن‌های موجود در همه هموارسازها که ما از آن‌ها استفاده می‌کنیم یک خصوصیت دارند و آن اینکه برای تمام t ها داریم

$$\sum_{i=1}^{n} W_i(t) = 1.$$

این نشان می‌دهد که هموارساز، منحنی را ثابت نگه می‌دارد. بنابراین اگر $Y_i = C$ برای تمام i ها باشد، $\hat{f}_n(t) = C$ خواهد بود. برای آگاهی بیشتر بغلانی (۱۳۹۴) را ببینید.

مثال ۱.۳.۳. (رگرسوگرام) فرض کنید $a < t_i < b \quad i = 1, 2, \ldots, n$ را در m بلوک مساوی تقسیم می‌کنیم که با $B_1, B_2, \ldots, B_m$ نمایش داده می‌شوند. $\hat{f}_n(t)$ را به صورت زیر تعریف می‌کنیم

$$\hat{f}_n(t) = \frac{1}{k_j} \sum_{i:t_i \in B_j} y_i, \qquad t_i \in B_j$$

به طوری که k_j تعداد نقاط موجود در B_j است. به عبارت دیگر برآورد $\hat{f}_n$ تابعی است که توسط میانگین‌گیری کردن y_i ها در هر بلوک به دست آمده است. این برآورد، رگرسوگرام[۹] گفته می‌شود.

اکنون تعریف می‌کنیم

$$W_i(t) = \begin{cases} \frac{1}{k_j} & \text{اگر} \quad t_i \in B_j \\ 0 & \text{اگر} \quad t_i \notin B_j \end{cases}$$

بنابراین خواهیم داشت $\hat{f}_n(t) = \sum_{i=1}^{n} y_i W_i(t)$. پس به نظر می‌رسد که بردار وزن‌های W_i این گونه باشد

$$\boldsymbol{W}^\top(t) = (0, 0, \ldots, \frac{1}{k_j}, \ldots, \frac{1}{k_j}, 0, \ldots, 0)$$

[۹] Regressogram

برای دیدن شکل ماتریس هموارساز $\boldsymbol{W}$ که $n = ۹$ و $m = ۳$ و $k_۱ = k_۲ = k_۳ = ۳$ باشد. همچنین ماتریس هموارساز $\boldsymbol{W}$ به شکل ماتریس مربع خواهد بود. بنابراین

$$\boldsymbol{W} = \frac{۱}{۳} \times \begin{pmatrix} ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & ۱ & ۱ \\ ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & ۱ & ۱ \\ ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & ۱ & ۱ \\ ۰ & ۰ & ۰ & ۱ & ۱ & ۱ & ۰ & ۰ & ۰ \\ ۰ & ۰ & ۰ & ۱ & ۱ & ۱ & ۰ & ۰ & ۰ \\ ۰ & ۰ & ۰ & ۱ & ۱ & ۱ & ۰ & ۰ & ۰ \\ ۱ & ۱ & ۱ & ۰ & ۰ & ۰ & ۰ & ۰ & ۰ \\ ۱ & ۱ & ۱ & ۰ & ۰ & ۰ & ۰ & ۰ & ۰ \\ ۱ & ۱ & ۱ & ۰ & ۰ & ۰ & ۰ & ۰ & ۰ \end{pmatrix}$$

۴.۳ هموارساز میانگین متحرک

در این بخش به معرفی یک هموارساز بسیار ساده به نام میانگین متحرک یا «میانگین پویا» می‌پردازیم. فرض کنید داده‌های $(t_۱, y_۱), \ldots, (t_n, y_n)$ در مدل

$$\boldsymbol{y} = f(\boldsymbol{t}) + \boldsymbol{\epsilon}$$

به ما داده شده باشد. فرض می‌کنیم t اسکالر بوده و داده‌ها به صورت $t_۱ \leq \cdots \leq t_n$ باشند. به طوری که t_i ها دارای فاصله واحد باشند. برآورد تابع $f(.)$ در t_i به صورت میانگین k مشاهده متوالی مرکزی در t_i است. (برای جلوگیری از ابهام بهتر است که k را فرد بگیریم). به صورت فرمولی تعریف می‌کنیم

$$\hat{f}(t_i) = \frac{۱}{k} \sum_{j=\underline{i}}^{\bar{i}} y_j \qquad (۴.۳)$$

به طوری که $\underline{i} = i - (k-۱)/۲$ و $\bar{i} = i + (k-۱)/۲$ نشان‌دهنده حد پایین و حد بالای علامت مجموع است. همچنین برآوردگر برابر است با

$$\hat{f}(t_i) = \frac{۱}{k} \sum_{j=\underline{i}}^{\bar{i}} \hat{f}(t_j) + \frac{۱}{k} \sum_{j=\underline{i}}^{\bar{i}} \epsilon_j \qquad (۵.۳)$$

اگر k (تعداد همسایگی‌های میانگین‌گیری شده) با افزایش n افزایش یابد سپس مطابق قضایای حد مرکزی، قسمت دوم در سمت راست به طور تقریبی نرمال با میانگین صفر و واریانس $\sigma_\epsilon^۲/k$ خواهد بود. برای آگاهی بیشتر روزبه (۱۳۹۰) را ببینید.

اغلب نوشتن هموارساز میانگین متحرک (وسایر هموارسازها) به صورت ماتریسی مناسب خواهد بود. S را ماتریس هموارساز قرار می‌دهیم که به صورت زیر تعریف می‌شود

$$\boldsymbol{S}_{(n-k+1)} = \begin{pmatrix} \frac{1}{k} & \dots & \frac{1}{k} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{k} & \dots & \frac{1}{k} & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & \frac{1}{k} & \dots & \frac{1}{k} & 0 \\ 0 & \dots & 0 & 0 & \frac{1}{k} & \dots & \frac{1}{k} \end{pmatrix}$$

بنابراین می‌توان رابطه (۱.۳) را به شکل بردار ـ ماتریس به صورت زیر نوشت

$$\hat{\boldsymbol{y}} = \hat{f}(\boldsymbol{t}) = \boldsymbol{S}\boldsymbol{y} \qquad (۶.۳)$$

به طوری که $\boldsymbol{t}$، $\boldsymbol{Y}$، $\hat{\boldsymbol{Y}}$ و $\hat{f}(\boldsymbol{t})$ بردار هستند.

۵.۳ هموارسازهای کرنل

اکنون برآوردگر کرنل (هسته) f را در نقطه t_0 به صورت زیر در نظر بگیرید

$$\hat{f}(t_0) = \sum_{i=1}^{n} w_i(t_0) y_i \qquad (۷.۳)$$

در اینجا تابع رگرسیون را در نقطه t_0 به عنوان یک وزن از مجموع y_i ها برآورد می‌کنیم به طوری که وزن‌های $w_i(t_0)$ به t_0 بستگی دارد.
یک راه مناسب برای ساخت مفهوم متوسط وزنی، به کار بردن یک تابع کیفی مرکزی در صفر است که در هر دو جهت با یک نرخ کاهش می‌یابد و به‌وسیله یک پارامتر مقیاس کنترل می‌شود. این توابع عموما به عنوان کرنل شناخته می‌شوند و توابع چگالی احتمال هستند. $k(.)$ را یک تابع کراندار قرار می‌دهیم که انتگرال آن ۱ و در اطراف ۰ متقارن است. وزن‌ها را به صورت زیر تعریف می‌کنیم

$$w_i(t_0) = \frac{\frac{1}{hn} k\left(\frac{t_i - t_0}{h}\right)}{\frac{1}{hn} \sum_{i=1}^{n} k\left(\frac{t_i - t_0}{h}\right)} \qquad (۸.۳)$$

شکل وزن‌ها توسط $k(.)$ تعیین می‌گردد و بزرگی مقدار آن‌ها توسط h کنترل می‌شود که به عنوان پهنای باند شناخته می‌شود. انتخاب پهنای باند h، مقدار هموواری را کنترل می‌کند. پهنای باند کم برآورد بسیار ناهنجاری می‌دهد، در حالی که پهنای باند بزرگتر

برآورد هموارتری به‌دست می‌آورد. هر چه مشاهدات از نقطه t_0 دور تر شوند، پارامتر h وزن بزرگتری به آن‌ها در برآورد $f(t_0)$ می‌دهد. با استفاده از رابطه (۸.۳) برآوردگر تابع رگرسیون ناپارامتری، در ابتدا توسط نادارای[۱۰] (۱۹۶۴) و واتسون[۱۱] (۱۹۶۴)، پیشنهاد گردید که به صورت زیر آمده است

$$\hat{f}(t_0) = \frac{\frac{1}{hn}\sum_{i=1}^{n} k\left(\frac{t_i - t_0}{h}\right) y_i}{\frac{1}{hn}\sum_{i=1}^{n} k\left(\frac{t_i - t_0}{h}\right)}. \qquad (۹.۳)$$

عموما انتخاب کرنل از انتخاب پهنای باند اهمیت کمتری دارد. ساده‌ترین کرنل، کرنل یکنواخت است (همچنین به کرنل مستطیلی و جعبه‌ای نیز می‌توان اشاره کرد)، به طوری که مقداری به اندازه $\frac{1}{2}$ در $[-1, 1]$ و در جاهای دیگر 0 به‌دست می‌دهد. اما کرنل نرمال و سایر کرنل‌ها نیز به طور گسترده مورد استفاده قرار می‌گیرند. (به وند و جونز[۱۲] ۱۹۹۵، برای مشاهده انواع بیشتری از هموارسازهای کرنل مراجعه شود).
در بسیاری از موارد به طور رایج از میانگین متحرک استفاده می‌شود. کرنل یکنواخت به آسانی میانگین مشاهداتی که در محدوده $t_0 \pm h$ قرار می‌گیرند را به‌دست می‌دهد.

۶.۳ مدل‌های رگرسیون نیمه‌پارامتری

رگرسیون نیمه‌پارامتری می‌تواند به عنوان یک نمونه دیگری از رگرسیون پارامتری و همچنین رفتار دو مدل پارامتری و ناپارامتری باهم، مورد بحث قرار گیرد. داشتن هر دو اجزای پارامتری و ناپارامتری در مدل آن را به یک مدل نیمه‌پارامتری تبدیل می‌کند.
مجموعه داده‌های $(y_1, \boldsymbol{x}_1, t_1), \ldots, (y_n, \boldsymbol{x}_n, t_n)$ از مدل رگرسیون نیمه پارامتری تبعیت می‌کند، هرگاه

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + f(t_i) + \epsilon_i, \qquad i = 1, \ldots, n \qquad (۱۰.۳)$$

به طوری که $\boldsymbol{x}_i^\top = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ip})$ برای $i = 1, 2, \ldots, n$ و $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ یک بردار از پارامترهای نامعلوم است. به طوری که $f(.)$ یک تابع مجهول و هموار است و t ها عضو یک مجموعه فشرده و به صورت $t_1 \leq t_2 \leq \cdots \leq t_n$ مرتب شده‌اند به گونه‌ای که n تعداد مشاهدات در نمونه است. فرض می‌کنیم $\boldsymbol{\epsilon}$ یک بردار n تایی با توزیع نامشخص که دارای $E(\epsilon) = 0$ و $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2 \boldsymbol{V}$ می‌باشد. مدل‌های رگرسیون

[۱۰] Nadaraya
[۱۱] Watson
[۱۲] Wand and Jones

نیمه‌پارامتری نسبت به مدل‌های خطی استاندارد انعطاف‌پذیرترند، به این دلیل که آن‌ها هم دارای قسمت پارامتری و هم ناپارامتری هستند. در واقع استفاده از این مدل زمانی مفید است که متغیر وابسته y به طور خطی با متغیر مستقل x و به طور غیر خطی با متغیر مستقل t رابطه داشته باشد.

۷.۳ برآورد به روش تفاضلی

در برآورد تفاضلی، ابتدا روندی را که توسط تابع $f(\cdot)$ در داده‌ها به وجود می‌آید را حذف نموده و سپس به برآورد پارامترهای مدل می‌پردازیم. بنابراین در این روش نیازی به برآورد تابع $f(\cdot)$ نداشته و به طور مستقیم پارامترها را برآورد می‌کنیم.
به طور خلاصه، با استفاده از روش تفاضلی می‌توان:

(۱) بدون نیاز به استفاده از تکنیک‌های برازش ناپارامتری، واریانس مانده‌ها را برآورد کرد.

(۲) آزمون فرضیه در مورد خطی بودن تابع ناپارامتری را انجام داد.

(۳) قسمت خطی مدل‌های نیمه‌پارامتری را برآورد کرد.

(۴) آزمون فرضیه در مورد خطی بودن تابع ناپارامتری در مدل‌های نیمه‌پارامتری انجام داد.

(۵) می‌توان آزمون فرضیه برابر بودن دو تابع در دو مدل مختلف در مدل‌های خطی نیمه‌پارامتری را انجام داد.

برای به‌دست آوردن شکلی از برآوردگر تفاضلی ابتدا مدل (۱۰.۳) را به صورت ماتریس و بردار به صورت زیر بازنویسی می‌کنیم

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + f(\boldsymbol{t}) + \boldsymbol{\epsilon} \qquad (۱۱.۳)$$

به طوری که $\boldsymbol{X} = (\boldsymbol{x}_۱, \ldots, \boldsymbol{x}_n)^\top$ و $\boldsymbol{x}_i^\top = (\boldsymbol{x}_{i۱}, \boldsymbol{x}_{i۲}, \ldots, \boldsymbol{x}_{ip}), i = ۱, ۲, \ldots, n$ که $\boldsymbol{X}$ یک ماتریس $n \times p$ می‌باشد و داشته باشیم $f(\boldsymbol{t}) = \big(f(t_۱), \ldots, f(t_n)\big)^\top$، $\boldsymbol{y} = (y_۱, \ldots, y_n)^\top$ و $\boldsymbol{\epsilon} = (\epsilon_۱, \ldots, \epsilon_n)^\top$. $\boldsymbol{d} = (d_۰, \ldots, d_m)$ را یک بردار $m+۱$ تایی در نظر می‌گیریم، به گونه‌ای که m مرتبه تفاضل و $d_۰, \ldots, d_m$ وزن‌های تفاضل با شرایط زیر باشند

$$\sum_{j=۰}^{m} d_j = ۰ \qquad , \sum_{j=۰}^{m} d_j^۲ = ۱ \qquad (۱۲.۳)$$

اکنون ماتریس تفاضلی $\boldsymbol{D}$ با ابعاد $(n-m)\times n$ که عناصر آن از رابطه (۱۲.۳) تبعیت می‌کنند را به صورت زیر تعریف می‌کنیم

$$\boldsymbol{D}=\begin{pmatrix} d_0 & d_1 & \dots & d_m & 0 & 0 & \dots & 0\\ 0 & d_0 & d_1 & \dots & d_m & 0 & \dots & 0\\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots\\ 0 & 0 & \dots & 0 & d_0 & d_1 & \dots & d_m \end{pmatrix}$$

به عنوان مثال اگر وزن‌ها را بصورت زیر تعریف کنیم، آن‌گاه شرایط (۱۲.۳) برقرار است:

$$d_0=\sqrt{\frac{m}{m+1}},=d_1=d_2=\cdots,d_m=-\sqrt{\frac{1}{m(m+1)}}$$

هال و همکارانش (۱۹۹۰) مقدار بهینه ضرایب تفاضلی را با استفاده از روش‌های عددی تا مرتبه $m=10$ به‌دست آوردند. این ضرایب در جدول (۱.۳) آورده شده‌اند. در عین حال روش مشخصی برای تعیین مقدار m پیشنهاد نشده است . برای آگاهی بیشتر روزبه (۱۳۹۴) را ببینید.

با استفاده از ماتریس تفاضلی در مدل (۱۱.۳) اجازه برآورد مستقیم اثر پارامتر را می‌دهد. در مجموع داریم

$$\boldsymbol{Dy}=\boldsymbol{DX\beta}+\boldsymbol{D}f(\boldsymbol{t})+\boldsymbol{D\epsilon} \qquad (۱۳.۳)$$

ماتریس تفاضلی $\boldsymbol{D}$ در مدل (۱۳.۳) اثر تابع ناپارامتری را در نمونه‌های بزرگ از بین می‌برد. بنابراین مدل (۱۳.۳) را می‌توان به شکل زیر باز نویسی نمود

$$\boldsymbol{Dy}=\boldsymbol{DX\beta}+\boldsymbol{D\epsilon}$$

یا

$$\tilde{\boldsymbol{y}}=\tilde{\boldsymbol{X}}\boldsymbol{\beta}+\tilde{\boldsymbol{\epsilon}} \qquad (۱۴.۳)$$

به گونه‌ای که $\tilde{\boldsymbol{y}}=\boldsymbol{Dy}$، $\tilde{\boldsymbol{X}}=\boldsymbol{DX}$ و $\tilde{\boldsymbol{\epsilon}}=\boldsymbol{D\epsilon}$. بنابراین، $\tilde{\boldsymbol{\epsilon}}$ یک بردار $(n-m)$ تایی دارای توزیع نامعلوم با $E(\tilde{\boldsymbol{\epsilon}})$ و $E(\tilde{\boldsymbol{\epsilon}}\tilde{\boldsymbol{\epsilon}}^{\top})=\sigma^2\boldsymbol{V}_D$ است به طوری که داریم $\boldsymbol{V}_D=\boldsymbol{DVD}^{\top}\neq I_{n-m}$ برآوردگر تفاضلی کمترین توان‌های دوم تعمیم‌یافته[۱۳] (DGLSE) به صورت زیر به دست می‌آید:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GD}^{LS} &= \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\beta})^{\top}\boldsymbol{V}_D^{-1}(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\beta})\\ &= \boldsymbol{C}_D^{-1}\tilde{\boldsymbol{X}}^{\top}\boldsymbol{V}_D^{-1}\tilde{\boldsymbol{y}} \qquad (۱۵.۳)\end{aligned}$$

به گونه‌ای که $\boldsymbol{C}_D=\tilde{\boldsymbol{X}}^{\top}\boldsymbol{V}_D^{-1}\tilde{\boldsymbol{X}}$.

[۱۳] Difference-based generalized least squares estimator

جدول ۱.۳: ضرایب تفاضلی بهینه

m	۱	۲	۳	۴	۵
$d_۰$	۰/۷۰۷۱	۰/۸۰۹۰	۰/۸۵۸۲	۰/۸۸۷۳	۰/۹۰۶۴
$d_۱$	−۰/۷۰۷۱	−۰/۵۰۰۰	−۰/۳۸۳۲	−۰/۳۰۹۹	−۰/۲۶۰۰
$d_۲$		−۰/۳۰۹۰	−۰/۲۸۰۹	−۰/۲۴۶۴	−۰/۲۱۹۷
$d_۳$			−۰/۱۹۴۲	−۰/۱۹۰۱	−۰/۱۷۷۴
$d_۴$				−۰/۱۴۰۹	−۰/۱۴۲۰
$d_۵$					−۰/۱۱۰۳
m	۶	۷	۸	۹	۱۰
$d_۰$	۰/۹۲۰۰	۰/۹۳۰۲	۰/۹۳۸۰	۰/۹۴۴۳	۰/۹۴۹۴
$d_۱$	−۰/۲۲۳۸	−۰/۱۹۶۵	−۰/۱۷۵۱	−۰/۱۵۷۸	−۰/۱۴۳۷
$d_۲$	−۰/۱۹۲۵	−۰/۱۷۲۸	−۰/۱۵۶۵	−۰/۱۴۲۹	−۰/۱۳۱۴
$d_۳$	−۰/۱۶۳۵	−۰/۱۵۰۶	−۰/۱۳۸۹	−۰/۱۲۸۷	−۰/۱۱۹۷
$d_۴$	−۰/۱۳۶۹	−۰/۱۲۹۹	−۰/۱۲۲۴	−۰/۱۱۵۲	−۰/۱۰۸۵
$d_۵$	−۰/۱۱۲۶	−۰/۱۱۰۷	−۰/۱۰۶۹	−۰/۱۰۲۵	−۰/۰۹۷۸
$d_۶$	−۰/۰۹۰۶	−۰/۰۹۳۰	−۰/۰۹۲۵	−۰/۰۹۰۵	−۰/۰۸۷۷
$d_۷$		−۰/۰۷۶۸	−۰/۰۷۹۱	−۰/۰۷۹۲	−۰/۰۷۸۲
$d_۸$			−۰/۰۶۶۶	−۰/۰۶۸۷	−۰/۰۶۹۱
$d_۹$				−۰/۰۵۸۸	−۰/۰۶۰۶
$d_{۱۰}$					−۰/۰۵۲۷

۸.۳ برآوردگر لاسوی مشتق‌پذیر تفاضلی

در این بخش برآوردگر لاسوی مشتق‌پذیر تفاضلی[۱۴](DDLASSO) را معرفی کرده و خواص آن را بررسی می‌کنیم. تحت مفروضات مدل (۱۴.۳) برآوردگر DDLASSO به صورت زیر به دست می‌آید:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{DLASSO} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}{\operatorname{argmin}}\left\{(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\beta})^{\top}\boldsymbol{V}_D^{-1}(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}P(\boldsymbol{\beta}_j,s)\right\}\\ &= \boldsymbol{C}_D^{-1}\tilde{\boldsymbol{X}}^{\top}\boldsymbol{V}_D^{-1}\tilde{\boldsymbol{y}}+\lambda\sum_{j=1}^{p}P(\boldsymbol{\beta}_j,s). \qquad (۱۶.۳)\end{aligned}$$

[۱۴] Differenced-based DLASSO

قضیه زیر توزیع مجانبی برآوردگر $\hat{\boldsymbol{\beta}}^{DLASSO}$ را ارائه می‌کنیم. از آن‌جایی که اثبات این قضیه مشابه اثبات قضیه ۲.۳.۲ است از آوردن جزئیات آن خودداری می‌کنیم.

قضیه ۱.۸.۳. فرض کنید $\boldsymbol{u} \in \mathbb{R}^p$، $S_n^* = \frac{s}{(n-m)^{\frac{1}{2}+\epsilon}} \to 0$، $\epsilon > 0$، $\frac{\lambda_n}{\sqrt{n-m}} \to \lambda_0 \geq 0$، $\frac{\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}}{n-m} \to \boldsymbol{\Sigma}^*$. در این صورت

$$\sqrt{n-m}(\hat{\boldsymbol{\beta}}^{DDLASSO} - \boldsymbol{\beta}) \to \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} k^*(\boldsymbol{u})$$

که در آن

$$k^*(\boldsymbol{u}) = -2\boldsymbol{u}^\top N(0, r\boldsymbol{\Sigma}^*) + \boldsymbol{u}^\top \boldsymbol{\Sigma}^* \boldsymbol{u} + \lambda_0 \sum_{i=1}^{p} \Big\{ u_i \text{Sign}(\beta_i) \mathbf{I}(\beta_i \neq 0) + |u_i| \mathbf{I}(\beta_i = 0) \Big\}.$$

بدیهی است الگوریتم دست‌یابی به برآوردگر $\hat{\boldsymbol{\beta}}^{DDLASSO}$ مشابه بخش ۴.۲ بوده با این تفاوت که $\boldsymbol{X}$ به $\tilde{\boldsymbol{X}}$ تبدیل می‌شود.

۹.۳ شبیه‌سازی

در این قسمت، در قالب یک مثال شبیه‌سازی برآوردگر لاسوی مشتق‌پذیر را در مدل تفاضلی به‌دست می‌آوریم. در این مطالعه، متغیر پاسخ را از مدل زیر به حجم $n = 100$ با 10^3 تکرار تولید می‌کنیم

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{5i}\beta_5 + f(t_i) + \epsilon_i, i = 1, \ldots, n \quad (۱۷.۳)$$

که در آن

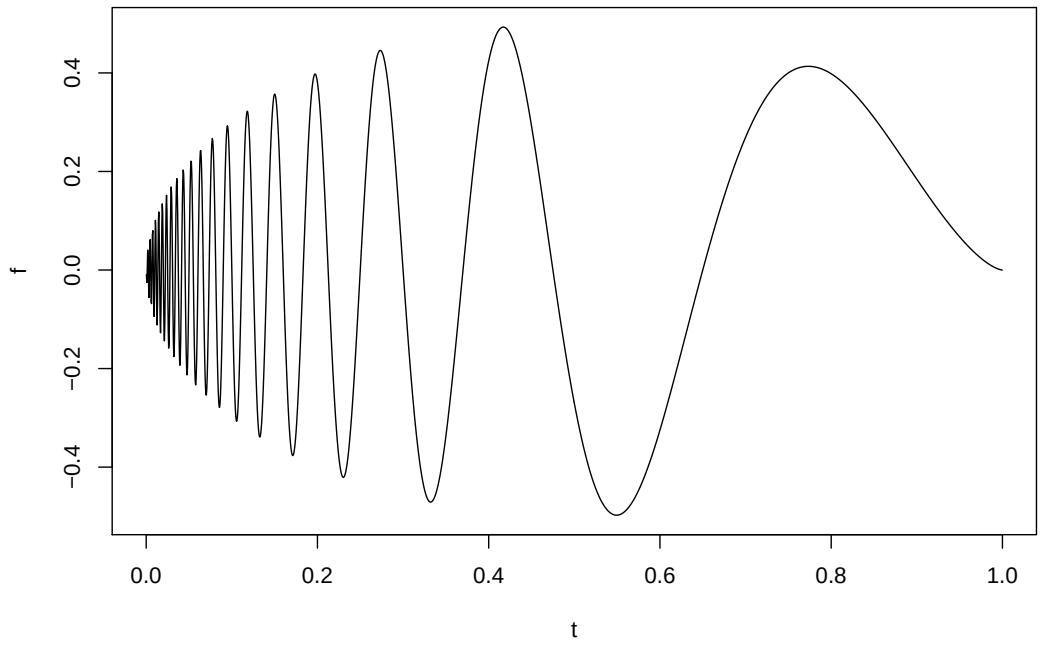
$$\boldsymbol{\beta} = (1/5, 2, 3, -5, 4)^\top$$

$$\boldsymbol{x}_i \sim N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma}_x), \boldsymbol{\mu} = \begin{pmatrix} -2 \\ 1 \\ 2 \\ 3 \\ 2 \end{pmatrix}, \boldsymbol{\Sigma}_x = \begin{pmatrix} 0/81 & 0/4 & 0/3 & -0/2 & -0/1 \\ 0/4 & 2/25 & 0/4 & 0/3 & -0/2 \\ 0/3 & 0/4 & 1 & 0/4 & 0/3 \\ -0/2 & 0/3 & 0/4 & 0/64 & 0/4 \\ -0/1 & -0/2 & 0/3 & 0/4 & 0/49 \end{pmatrix}$$

$$f(t_i) = \sqrt{t_i(1-t_i)} \sin\left(\frac{2/1\pi}{t_i + 0/05}\right), t_i = (i - 0/5)/n$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2 \boldsymbol{V}), \sigma_\epsilon^2 = 0/01, v_{ij} = (\frac{1}{n})^{|i-j|+1}$$

تابع ناپارامتری در نظر گرفته شده، تابع داپلر[۱۵] نامیده می‌شود. تابع داپلر به دلیل شکل موجی و ناهمگن بودنش (شکل ۱.۳) به سختی برآورد می‌شود و به همین دلیل استفاده از آن برای نشان دادن کارایی روش تفاضلی و روش هموارسازی به‌کار رفته می‌باشد.



شکل ۱.۳: تابع داپلر (تابع ناپارامتری مدل در نظر گرفته شده) به ازای $n = ۲۰۰۰$

شبیه‌سازی مونت کارلو به ازای $N = ۱۰۰۰$ انجام شد و برآوردگرهای پیشنهادی به دست آورده شد. قسمت پارامتری مدل (۱۷.۳)، به عبارتی β با استفاده از ضرایب تفاضلی مرتبه چهارم، $d_۰ = ۰/۸۸۷۳$، $d_۱ = -۰/۳۰۹۹$، $d_۲ = -۰/۲۴۶۵$، $d_۳ = -۰/۱۹۰۱$ و $d_۴ = -۰/۱۴۰۹$ برآورد می‌شود. آن‌گاه با جایگزینی برآورد به‌دست آمده برای پارامتر β، تابع ناپارامتری $f(\cdot)$ را با استفاده از تکنیک هموارسازی کرنل و بکار بردن کرنل نرمال، برآورد می‌کنیم. لذا ماتریس تفاضلی از مرتبه $۱۰۰ \times (۴ - ۱۰۰)$ بصورت زیر

[۱۵] Doppler function

خواهد بود

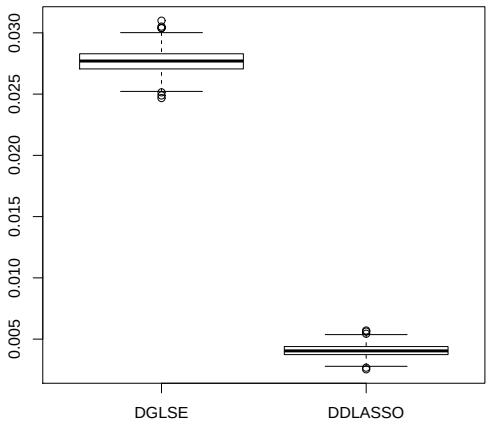
$$\boldsymbol{D} = \begin{pmatrix} d_0 & d_1 & d_2 & d_3 & d_4 & 0 & 0 & \dots & 0 \\ 0 & d_0 & d_1 & d_2 & d_3 & d_4 & 0 & \dots & 0 \\ \vdots & \ddots & & & & & & \vdots & \\ 0 & 0 & \dots & 0 & d_0 & d_1 & d_2 & d_3 & d_4 \end{pmatrix}$$

مخاطره برآوردگرهای پیشنهادی به‌صورت زیر به دست می‌آید.

$$\mathrm{R}(\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}\right)^{\top} \left(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}\right),$$

که $\hat{\boldsymbol{\beta}}^{(i)}$ برآوردگرهای به‌دست آمده در تکرار m ام می‌باشد.

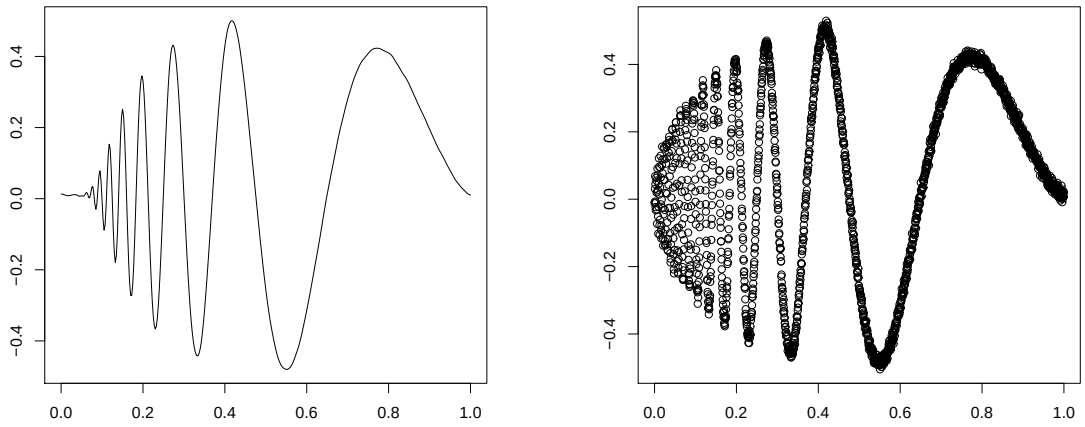
شکل ۲.۳ مقایسه مخاطره برآوردگر لاسوی مشتق‌پذیر تفاضلی را با برآوردگر تفاضلی نشان می‌دهد. بر اساس این نمودار مشخص می‌شود که استفاده از برآوردگر لاسوی مشتق‌پذیر در روش تفاضلی باعث کاهش مخاطره می‌گردد.



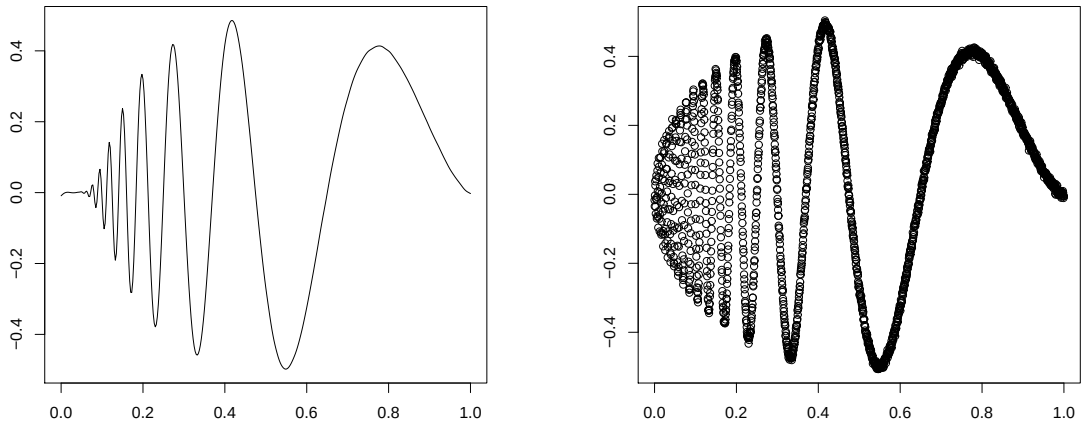
شکل ۲.۳: نمودار جعبه‌ای مقایسه مخاطره برآوردگرهای لاسوی مشتق‌پذیر تفاضلی و کمترین توان‌های دوم تعمیم‌یافته

در شکل ۳.۳ (قاب)، نمودار مانده‌های برآوردشده با روش هموارسازی کرنل پس از برآورد قسمت خطی به‌وسیله برآوردگر تفاضلی تعمیم یافته یعنی $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GD}^{LS}$ در مقابل t_i ها و منحنی ناپارامتری برازش‌شده به این مانده‌ها را نشان می‌دهد. نمودارهای قاب پایین با جایگزینی $\hat{\boldsymbol{\beta}}_{GD}^{LS}$ با $\hat{\boldsymbol{\beta}}^{DDLASSO}$ به دست آمده است.

$\hat{\boldsymbol{\beta}}_{GD}^{LS}$



$\hat{\boldsymbol{\beta}}^{DDLASSO}$



شکل ۳.۳: برآورد تابع داپلر به روش هموار سازی کرنل، بالا سمت راست: مانده‌ها پس از براورد قسمت خطی با استفاده از $\hat{\boldsymbol{\beta}}_{GD}^{LS}$، بالا سمت چپ: برازش منحنی هموار به مانده‌ها با استفاده از روش هموارسازی کرنل و به‌کار بردن نرمال با پهنای باند بهینه $h = ۰/۰۰۵$، پایین مانند قسمت بالاست که در آن به جای $\hat{\boldsymbol{\beta}}_{GD}^{LS}$ از $\hat{\boldsymbol{\beta}}^{DDLASSO}$ استفاده شده است.

فصل ۴

تحلیل مثال‌های واقعی

۱.۴ مقدمه

در این فصل، در دو مجموعه داده واقعی، کاربردی از روش‌های بیان‌شده در این پایان‌نامه را بررسی می‌کنیم. ابتدا، در مجموعه داده‌های پروستات، کارایی روش لاسوی مشتق‌پذیر را با سایر برآوردگرها مورد مطالعه قرار می‌دهیم. سپس در مجموعه داده‌های مطالعه پنلی روی درآمدهای پویا[۱] PSID (امرز، ۱۹۸۷)، برآوردگر لاسوی مشتق‌پذیر تفاضلی را با برآوردگرهای لاسوی تفاضلی و نیز برآوردگر تفاضلی، مورد بررسی قرار می‌دهیم.

۲.۴ تحلیل داده‌های سرطان پروستات

مجموعه داده‌های سرطان پروستات، ارتباط بین سطح آنتی‌ژن اختصاصی پروستات (پادتن مخصوص پروستات) با اندازه‌های بالینی ریشه غده پروستات ۹۷ مرد، که اشعه رادیکال پرستاکتومی را دریافت کرده‌اند، و هشت متغیر زیر را بررسی می‌کند: لگاریتم حجم سرطان (lcanvol)، لگاریتم وزن پروستات (lweight)، لگاریتم مقدار یاخته

[۱] Panel study on income dynamics

پروستات خوش‌خیم (lbph)، لگاریتم خصوصیات کپسول نفوذی (lcp)، سن، نمره گلیسون (gleason)، درصد گلیسون ۴ یا ۵ (pgg45)، تهاجم بافتی هسته‌ای (svi). همه متغیرها استاندارد شده‌اند تا میانگین صفر و واریانس یک را داشته باشند و متغیر پاسخ مرکزی شده است که میانگین آن در این صورت صفر است. توصیفی از این داده‌ها در جدول ۱.۴ ارائه شده است.

جدول ۱.۴: جدول توصیف داده‌های سرطان پروستات

متغیرها	کمینه	چارک اول	میانه	میانگین	چارک سوم	بیشینه
lcanvol	−۱/۳۴۷۱	۰/۵۱۲۸	۱/۴۴۶۹	۱/۳۵۰۰	۲/۱۲۷۰	۳/۸۲۱۰
lweight	۲/۳۷۵	۳/۳۷۶	۳/۶۲۳	۳/۶۲۹	۳/۸۷۶	۴/۷۸۰
lbph	−۱/۳۸۶۳	−۱/۳۸۶۳	۰/۳۰۰۱	۰/۱۰۰۴	۱/۵۵۸۱	۲/۳۲۶۳
lcp	−۱/۳۸۶۳	−۱/۳۸۶۳	−۰/۷۹۸۵	−۰/۱۷۹۴	۱/۱۷۸۷	۲/۹۰۴۲
age	۴۱/۰۰	۶۰/۰۰	۶۵/۰۰	۶۳/۸۷	۶۸/۰۰	۷۹/۰۰
gleason	۶/۰۰۰	۶/۰۰۰	۷/۰۰۰	۶/۷۵۳	۷/۰۰۰	۹/۰۰۰
pgg45	۰/۰۰	۰/۰۰	۱۵/۰۰	۲۴/۳۸	۴۰/۰۰	۱۰۰/۰۰
svi	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۲۱۶۵	۰/۰۰۰۰	۱/۰۰۰۰
lpsa	−۰/۴۳۰۸	۱/۷۳۱۷	۲/۵۹۱۵	۲/۴۷۸۴	۳/۰۵۶۴	۵/۵۸۲۹

روش‌های لاسو، ریج، SCAD، LS، Enet و لاسوی مشتق‌پذیر روی داده‌ها به‌کار گرفته شده‌اند تا تصویر واضحی از شباهت برآوردگر لاسوی مشتق‌پذیر با سایر برآوردگرهای جریمه‌شده به‌دست آید. پارامترهای تنظیم‌کننده با BIC انتخاب شده‌اند. مدل‌ها براساس BIC و AIC مقایسه شده‌اند. تعریف AIC و BIC را در پیوست آ.۱.۲ ببینید.

کارایی برآوردگر لاسوی مشتق‌پذیر به ازای مقدار ثابت $s = ۰/۰۰۱$ با برآوردگرهای لاسو، Enet و SCAD در جدول ۲.۴ مقایسه شده‌اند. برآوردگرهای لاسوی مشتق‌پذیر، لاسو و Enet، هر سه، پنج متغیر انتخاب می‌کنند و بر اساس معیارهای AIC و BIC بر برآوردگر SCAD برتری دارند. جدول ۳.۴ مقدار ضرایب این برآوردگرها را نشان می‌دهد.

اگر برای لاسوی مشتق‌پذیر، s را ثابت و برابر ۱ قرار دهیم، انتظار داریم نتیجه شبیه به برآوردگر ریج باشد. جدول ۴.۴ این ادعا را تایید می‌کند. لازم به ذکر است که هدف از ارائه مقادیر AIC و BIC انتخاب مدل و مقایسه دو روش لاسوی مشتق‌پذیر و ریج نبوده، بلکه هدف این است که نشان دهیم چنانچه پارامتر s مقدار یک را انتخاب کند نتایج لاسوی مشتق‌پذیر مشابه نتایج ریج است. همانطور که از مقادیر AIC و BIC

جدول ۲.۴: مقایسه برآوردگرهای لاسوی مشتق‌پذیر، لاسو، Enet، SCAD با $s =$ ۰/۰۰۱، روی مجموعه داده‌های پروستات.

روش	دقت	AIC	BIC	درجه آزادی	متغیرهای مهم
لاسوی مشتق پذیر	$s =$ ۰/۰۰۱	۲۰۷/۶	۲۱۶/۱	۵	lcavol، lbph، lweight، pgg45، svi
لاسو	–	۲۰۶/۷	۲۱۵/۳	۵	lcavol، lbph، lweight، pgg45، svi
Enet	$\alpha =$ ۰/۰۰۱	۲۰۶/۸	۲۱۵/۳	۵	lcavol، lbph، lweight، pgg45، svi
SCAD	–	۲۱۴/۶	۲۳۱	۴	lcavol، lbph، lweight، svi

جدول ۳.۴: مقادیر ضرایب برآوردگرهای لاسوی مشتق‌پذیر ($s =$ ۰/۰۰۱)، لاسو، Enet، SCAD روی مجموعه داده‌های پروستات.

	روش			
	لاسوی مشتق‌پذیر	لاسو	Enet	SCAD
lcavol	۰/۵۲۷۳	۰/۵۹۵۵	۰/۵۲۴۶	۰/۶۰۴۱
lweight	۰/۱۴۹۳	۰/۱۴۴۲	۰/۱۴۵۷	۰/۱۹۲۹
age	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
lbph	۰/۰۶۴۸	۰/۱۳۲۶	۰/۰۵۹۳	۰/۱۳۵۲
svi	۰/۲۰۳۱	۰/۲۷۰۹	۰/۱۹۹۹	۰/۲۷۱۷
lcp	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
gleason	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
pgg45	۰/۰۳۷۱	۰/۱۰۶۹	۰/۰۳۴۰	۰/۰۰۰۰

نتیجه می‌شود به ازای $s = ۱$ لاسوی مشتق‌پذیر مشابه ریج عمل می‌کند. در جدول ۵.۴ مقدار ضرایب آمده‌اند.

جدول ۴.۴: مقایسه برآوردگر لاسوی مشتق‌پذیر و ریج

روش	دقت	AIC	BIC	درجه آزادی	متغیرهای مهم
لاسوی مشتق پذیر	۱	۲۰۷/۴۴	۲۲۷	۸	همه متغیرها
ریج	–	۲۰۷/۱	۲۲۶	۸	همه متغیرها

وقتی در برآوردگر لاسوی مشتق‌پذیر، s را ثابت و برابر ۱۰۰ قرار دهیم انتظار داریم که پاسخ شبیه برآوردگر LS باشد. جدول ۶.۴ انتظار ما را تایید می‌کند و توضیحات مربوط به مقادیر AIC و BIC جدول ۴.۴ در این‌جا نیز صادق است. جدول ۷.۴ مقدار ضرایب را نشان می‌دهد.

مقدار BIC در جدول ۷.۴ نشان می‌دهد که لاسوی مشتق‌پذیر بهتر از برآوردگر LS است، که این موضوع احتمالا به دلیل تاثیر مقدار کمی پارامتر تنظیم‌کننده برای $s = ۱۰۰$ است.

جدول ۵.۴: مقادیر ضرایب برآوردگرهای لاسوی مشتق‌پذیر ($s = ۱$) و ریج روی مجموعه داده‌های پروستات.

	روش	
	لاسوی مشتق‌پذیر	ریج
lcavol	۰/۴۶۵۲	۰/۴۶۶۴
lweight	۰/۱۷۹۷	۰/۱۸۹۷
age	−۰/۰۷۲۷	−۰/۰۹۵
lbph	۰/۱۰۵۷	۰/۱۱۸۴
svi	۰/۲۲۲۹	۰/۲۴۵۶
lcp	−۰/۰۰۸۵	−۰/۰۰۴۰
gleason	۰/۰۲۸۶	۰/۰۳۹۰
pgg45	۰/۰۷۱۲	۰/۰۸۲۹

جدول ۶.۴: مقایسه برآوردگر لاسوی مشتق‌پذیر و LS

روش	دقت	AIC	BIC	درجه آزادی	متغیرهای مهم
لاسوی مشتق‌پذیر	۱۰۰	۲۰۴	۲۲۸	۸	همه متغیرها
LS	–	۲۰۲	۲۳۳	۸	همه متغیرها

جدول ۷.۴: مقادیر ضرایب برآوردگرهای لاسوی مشتق‌پذیر ($s = ۱۰۰$) و LS روی مجموعه داده‌های پروستات.

	روش	
	لاسوی مشتق‌پذیر	LS
lcavol	۰/۴۵۱۰	۰/۵۹۹۳
lweight	۰/۱۸۰۹	۰/۱۹۵۵
age	−۰/۰۷۱۴	−۰/۱۲۶۵
lbph	۰/۱۰۵۲	۰/۱۳۴۵
svi	۰/۲۲۴۲	۰/۲۷۴۷
lcp	−۰/۰۰۱۹	−۰/۱۲۷۷
gleason	۰/۰۲۸۲	۰/۰۲۸۲
pgg45	۰/۰۷۰۹	۰/۱۱۰۵

۳.۴ مجموعه داده‌های PSID

امرز (۱۹۸۷) از نمونه‌ای در مورد مطالعه پنلی روی درآمدها پویا استفاده نمود تا به‌طور سیستماتیک چندین نظریه و فرضیه‌های آماری را که تا آن زمان در مورد مدل‌های تجربی درآمد زنان بررسی شده بود را مطالعه کند. مجموعه داده‌های PSID به‌طور رایگان از سایت https://ideas.repec.org/p/boc/bocins/mroz.html قابل دسترسی است.

این داده‌ها از بین زنان سفیدپوست متاهل آمریکایی بین ۳۰ تا ۶۰ ساله در سال ۱۹۷۵ جمع‌آوری شده است که ۷۵۳ مشاهده و ۱۹ متغیر به‌شرح زیر دارد:

- INLF (اگر برابر ۱ باشد، بدین معنی است که در سال ۱۹۷۵، نیروی کار اجباری بوده است)
- HOURS (ساعت‌های کاری در سال ۱۹۷۵)

- k5 (کودک زیر ۶ سال)
- k618 (کودکان ۶ تا ۱۸ سال)
- AGE (سن زنان بر حسب سال)
- EDUC (سال‌های تحصیل در مدرسه)
- WAGE (برآورد حقوق ساعتی از درآمد)
- REPWAGE (نرخ حقوق گزارش شده در سال ۱۹۷۶)
- HUSHRS (ساعات کاری شوهر در سال ۱۹۷۵)
- HUSAGE (سن شوهر)
- HUSEDUC (سال‌های تحصیل در مدرسه شوهر)
- HUSWAGE (حقوق ساعتی شوهر در سال ۱۹۷۵)
- FAMINC (درآمد خانواده در سال ۱۹۷۵)
- MTR (نرخ مالیات زنان)
- MOTHEDUC (سال‌های تحصیل در مدرسه مادران)
- FATHEDUC (سال‌های تحصیل در مدرسه پدران)
- UNEM (نرخ بیکاری در کشور)
- CITY (اگر در شهر SMSA زندگی کند، مقدار ۱ را می‌پذیرد)
- EXPER (تجربه بازار کار)
- NWIFEINC (برابر با (FAMINC - WAGE))/1000 است)

مشابه امرز (۱۹۸۷)، متغیر HOURS، ساعات کاری زنان در سال ۱۹۷۵ را به‌عنوان متغیر پاسخ در نظر می‌گیریم. مانند رحیم و همکاران (۲۰۱۲)، با توجه به ماهیت متغیر پاسخ، فقط بخشی از داده‌ها مربوط به زنان شاغل به کار اجباری را در نظر می‌گیریم. بنابراین ۲۴۸ مشاهده در مجموعه داده جدید داریم. مدل پیشنهادی شامل متغیرهای زیر است:

- AGE
- NWIFEINC
- K5
- k618
- WC (اگر مقدار متغیر EDUC از ۱۲ بیشتر باشد، مقدار ۱ می‌گیرد.)
- textschc (اگر مقدار متغیر HUSEDUC از ۱۲ بیشتر باشد، مقدار ۱ می‌گیرد.)
- UNEM
- EXPER
- MTR

آن‌ها مدل زیر را پیشنهاد کردند

$$hours = wc\beta_۱ + g(nwifeinc) + mtr\beta_۲ + exper\beta_۳ + unem\beta_۴ + k۵\beta_۵ + age\beta_۶ + k۶۱۸\beta_۷ + hc\beta_۸ + \epsilon \quad (۱.۴)$$

در اینجا متغیر NWIFEINC مولفه ناپارامتری است. هال و همکاران (۱۹۹۰) مقادیر بهینه ماتریس $\boldsymbol{D}$ را در روش تفاضلی با روش‌های عددی به‌دست آورند. در این جا از روش تفاضلی مرتبه $m = ۶$ استفاده می‌شود. بر اساس مقادیر پیشنهادی هال و همکاران (۱۹۹۰)،

$$\boldsymbol{d} = (۰ : ۰۹۰۶; ۰ : ۱۹۲۶; ۰ : ۱۳۶۹; ۰ : ۱۶۳۵; ۰ : ۱۹۲۵; ۰ : ۲۲۳۸; ۰ : ۹۲۰۰) \quad (۲.۴)$$

و در نتیجه ماتریس $\boldsymbol{D}$ ساخته می‌شود.

در ادامه، روش اعتبارسنجی متقابل K تایی برای به‌دست آوردن برآوردی از خطاهای پیش‌گویی مدل استفاده می‌شود. در این روش، مجموعه داده را به‌صورت تصادفی به K زیرمجموعه نسبتا مساوی تقسیم می‌کنیم. یکی از مجموعه‌ها را کنار می‌گذاریم، یعنی $\left\{\left(\boldsymbol{X}_D^{\text{test}}, \boldsymbol{y}_D^{\text{test}}\right)\right\}$، که آن را مجموعه آزمون می‌نامیم و از $K-۱$ مجموعه باقیمانده که آن‌ها را مجموعه آموزشی می‌نامیم برای برازش مدل استفاده می‌کنیم. برآوردگر به‌دست آمده را $\hat{\boldsymbol{\beta}}^{\text{train}}$ می‌نامیم. مدل برازش‌شده برای پیش‌گویی متغیر پاسخ در

مجموعه داده آزمون به‌کار می‌رود. سرانجام خطاهای پیش‌گویی بر اساس توان دوم خطای مقادیر مشاهده‌شده از مقادیر پیش‌گویی‌شده به‌دست می‌آید. یعنی

$$\mathrm{PE}^k = \left\| \boldsymbol{y}_{Dk}^{\text{test}} - \hat{\boldsymbol{y}}_{Dk}^{\text{test}} \right\|^2; \qquad k = 1, \ldots, K,$$

که در آن $\hat{\boldsymbol{y}}_{Dk}^{\text{test}} = \boldsymbol{X}_{Dk}^{\text{test}} \hat{\boldsymbol{\beta}}_{Dk}^{\text{train}}$.

برآیند برای همه K زیرمجموعه تکرار می‌شود و خطاهای پیش‌گویی محاسبه می‌شوند. برای اثر دادن تغییرات تصادفی اعتبارسنجی متقابل، فرآیند N بار تکرار می‌شود و متوسط خطای پیش‌گویی[۲] (APE) به‌صورت زیر محاسبه می‌شود

$$\mathrm{APE} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{K} \sum_{k=1}^{K} \mathrm{PE}_i^k \right),$$

که PE_i^k خطای پیش‌گویی k امین مجموعه آزمون در i امین تکرار است. روش اعتبارسنجی متقابل به طور مفصل‌تری در پیوست آ.۲.۱ آمده است.

کارایی برآوردگر دلخواه $\hat{\beta}^\star$ نسبت به برآوردگر لاسوی مشتق‌پذیر تفاضلی، $\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}}$، با فرمول کارایی[۳] (Eff) به‌صورت زیر به‌دست می‌آید.

$$\mathrm{Eff}(\hat{\boldsymbol{\beta}}^\star; \hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}}) = \frac{\mathrm{APE}(\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}})}{\mathrm{APE}(\hat{\boldsymbol{\beta}}^\star)}.$$

اگر مقدار کارایی از ۱ بیشتر باشد، آن‌گاه $\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}}$ بهتر از برآوردگر $\hat{\boldsymbol{\beta}}^\star$ رفتار می‌کند.

جدول ۸.۴ برآوردگرها و کارایی آن‌ها را به‌ازای $N = ۵۰۰۰$ نشان می‌دهد.

برای برآورد قسمت پارامتری، روش تفاضلی مرتبه ششم ($m = ۶$) به‌کار برده شد. اکنون از روش کرنل با پهنای $h_n = ۱/۲$ برای برآورد متغیر hours استفاده می‌شود. به منظور برآورد مولفه ناپارامتری، ابتدا پارامتر را با روش تفاضلی برآورد می‌کنیم و سپس رهیافت کرنل را برای برازش $Z_i(\hat{\boldsymbol{\beta}}^*) = \mathrm{hours}_i - \boldsymbol{x}_{Di}^\top \hat{\boldsymbol{\beta}}^*$ روی nwifeinc استفاده می‌کنیم که $\boldsymbol{x}_i = (\mathrm{exper}, \mathrm{mtr}, \mathrm{wc}, \mathrm{age}, \mathrm{unem}, \mathrm{k۵}, \mathrm{k۶۱۸}, \mathrm{hc})$ و $\hat{\boldsymbol{\beta}}^*$ یکی از برآوردگرهای تفاضلی، لاسوی تفاضلی و لاسوی مشتق‌پذیر تفاضلی است (شکل ۱.۴).

همان طور که در جدول ۸.۴ دیده می‌شود، با توجه به این که مقدار Eff برای لاسوی مشتق‌پذیر تفاضلی ۱/۱۲ است، این برآوردگر کاراتر از هر دو برآوردگر تفاضلی و لاسوی تفاضلی است. این نتیجه یکی از مهم‌ترین یافته‌های این پایان‌نامه می‌باشد که نشان می‌دهد در مدل نیمه‌پارامتری لاسوی مشتق‌پذیر تفاضلی برآوردگری کاراتر

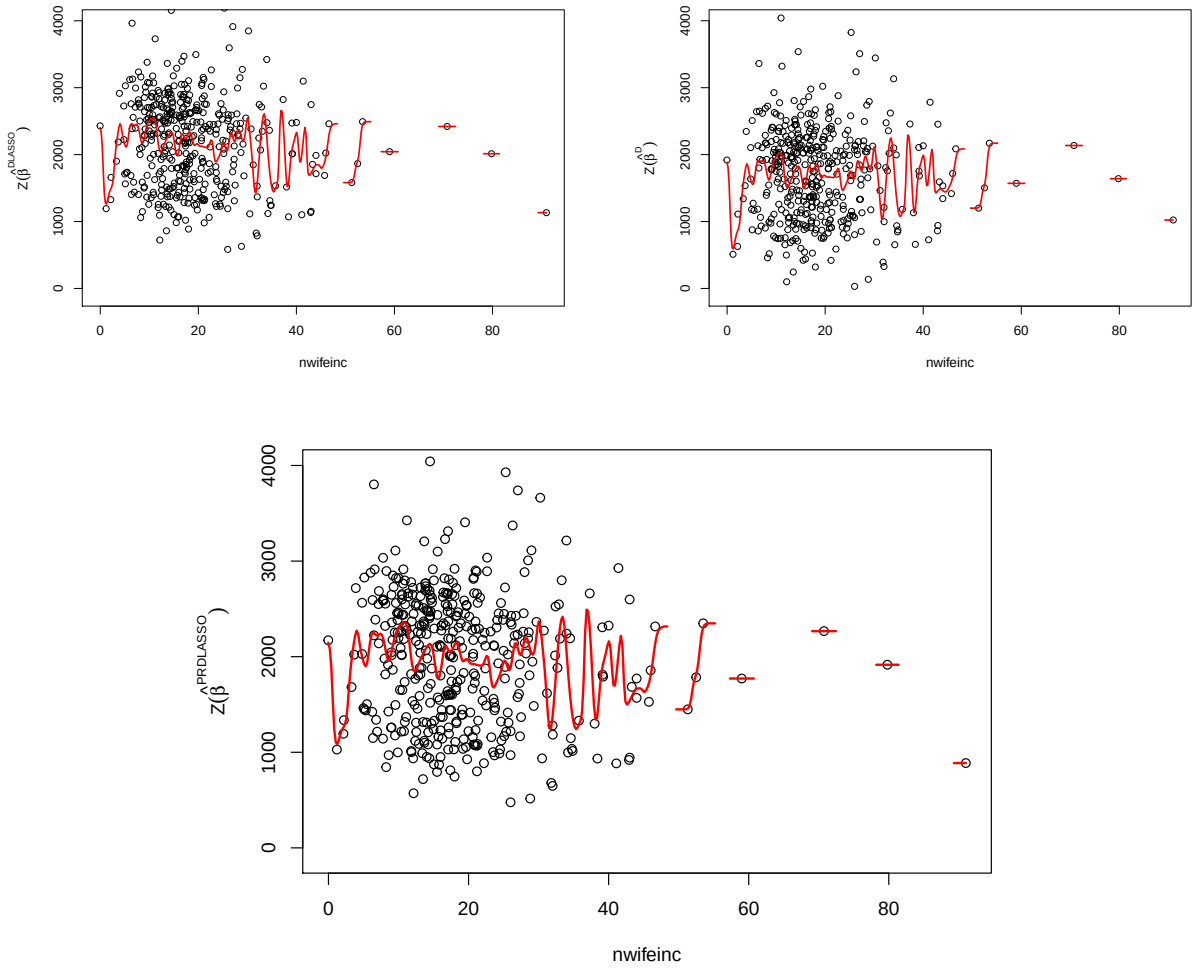
[۲] Average prediction error

[۳] Efficiency

جدول ۸.۴: مقادیر برآوردگرها و کارایی آن‌ها

متغیرها	تفاضلی	لاسوی تفاضلی	لاسوی مشتق‌پذیر تفاضلی
exper	۳۰/۲۶	۲۳/۵۷	۳۱/۸۷
mtr	−۴۱۶/۰۰	−۱۲۱۳/۵۰	−۷۰۷/۱۸
wc	−۶۸/۸۴	−۴۵/۵۶	−۱۱۰/۴۳
age	−۸/۳۶	−۵/۹۲	−۸/۷۵
unem	−۹/۸۵	−۹/۱۹	−۴/۱۰
k۵	−۱۲۸/۱	−۱۳۲/۱۱	−۱۳۹/۴۲۸۴
k۶۱۸	−۴۳/۰۶	−۲۵/۵۶	−۳۶/۱۸
hc	۲۲/۲۲	۰/۰۰	۲۴۴۳
Eff	۱/۰۰	۱/۰۵	۱/۱۲

نسبت به سایر برآوردگرهای جریمه‌شده برای انتخاب متغیر بر اساس متوسط خطای پیشگویی مدل است.



شکل ۱.۴: نمودار برآورد تابع ناپارامتری

پیوست آ

تعاریف و قضایای اساسی

آ.۰ مرکزی کردن مشاهدات

در این قسمت نشان می‌دهیم با مرکزی کردن مشاهدات در مدل رگرسیون خطی ساده، می‌توان اثر پارامتر عرض از مبدأ β_0 را حذف کرد. مدل (۱.۱) را در نظر بگیرید. با کم کردن میانگین مشاهدات از طرفین مدل رگرسیون خطی داریم

$$(y_i - \bar{y}) + \bar{y} = \beta_0 + \beta(x_i - \bar{x}) + \beta\bar{x} + \epsilon_i.$$

در نتیجه

$$(y_i - \bar{y}) = (\beta_0 + \beta\bar{x} - \bar{y}) + \beta(x_i - \bar{x}) + \epsilon_i. \qquad (آ.۱)$$

همچنین با میانگین‌گری از معادله (۱.۱) داریم

$$\bar{y} = \beta_0 + \beta\bar{x}. \qquad (آ.۲)$$

با جای‌گذاری معادله (آ.۲) در (آ.۱) نتیجه می‌شود

$$(y_i - \bar{y}) \quad = \quad (\beta_0 + \beta\bar{x} - \beta_0 - \beta\bar{x}) + \beta(x_i - \bar{x}) + \epsilon_i$$

$$= \beta(x_i - \bar{x}) + \epsilon_i.$$

فرض کنید $y_i^* = (y_i - \bar{y})$ و $x_i^* = (x_i - \bar{x})$. در این صورت مدل نهایی با رابطه زیر داده می‌شود

$$y_i^* = \beta x_i^* + \epsilon_i \qquad (۳.آ)$$

همانطور که در مدل (۳.آ) مشاهده می‌شود اثر $\beta_۰$ از مدل حذف شده است. به طور مشابه می‌توان در مدل رگرسیون چندگانه نیز اثر $\boldsymbol{\beta}_۰$ را حذف کرد.

تعریف آ.۱.۰ (نرم‌ـ L_p). فرض کنید $\mathbf{x} \in \mathbb{R}^۱$ یک بردار و $p \geq ۱$ یک عدد صحیح باشد. سپس

$$\|\mathbf{x}\|_p = \left(\sum_{i=۱}^{n} |x_i|^p \right)^{\frac{۱}{p}} \qquad (۴.آ)$$

نرم‌ـ L_p از بردار $\mathbf{x}$ نامیده می‌شود. واضح است

$$\|\mathbf{x}\|_p^p = \sum_{i=۱}^{n} |x_i|^p \qquad (۵.آ)$$

تعریف آ.۲.۰ (تابع خطا). تابع خطا به صورت زیر تعریف می‌شود

$$\text{erf} z = \frac{۲}{\sqrt{\pi}} \int_۰^z \exp\{-t\}^۲ dt$$

و اگر $z = \frac{x}{s}$

$$\text{erf}^c z = \frac{۲}{\sqrt{\pi}} \int_z^\infty \exp\{-t\}^۲ dt$$

از رامیز و سانچزا (۲۰۱۴) رابطه زیر را داریم

$$\frac{۱}{x + \sqrt{x^۲ + ۲}} < \exp\{\frac{x}{s}\}^۲ \int_{\frac{x}{s}}^\infty \exp\{-t\}^۲ dt \leq \frac{۱}{x + \sqrt{x^۲ + \frac{۴}{\pi}}}$$

$$\frac{۲}{\sqrt{\pi}} \frac{e^{-(\frac{x}{s})^۲}}{(\frac{x}{s}) + \sqrt{(\frac{x}{s})^۲ + ۲}} < \text{erf}^c\left(\frac{x}{s}\right) \leq \frac{۲}{\sqrt{\pi}} \frac{e^{-(\frac{x}{s})^۲}}{(\frac{x}{s}) + \sqrt{(\frac{x}{s})^۲ + \frac{۴}{\pi}}}$$

آ.۱ تابع محدب

فرض کنیم $-\infty \leq a \leq b \leq \infty$، تابع $f : (a, b) \to \mathbb{R}$ را محدب می‌گوییم در صورتی که به ازای هر دو عدد $x_۱, x_۲ \in (a, b)$ و هر t که $۰ \leq t \leq ۱$، داشته باشیم:

$$\forall x_۱, x_۲ \in X, \forall t \in [۰, ۱] : \qquad f(tx_۱ + (۱ - t)x_۲) \leq tf(x_۱) + (۱ - t)f(x_۲).$$

اگر در تعریف بالا تساوی را برداریم آن‌گاه f را اکیدا محدب می‌نامیم.

$$\forall x_۱ \neq x_۲ \in X, \forall t \in (۰, ۱): \qquad f(tx_۱ + (۱-t)x_۲) \leq tf(x_۱) + (۱-t)f(x_۲).$$

خاصیت‌ها:

$$R(x_۱, x_۲) = \frac{f(x_۱) - f(x_۲)}{x_۱ - x_۲} \tag{۶.آ}$$

$$\forall x_۱, x_۲ \in C : f\left(\frac{x_۱ + x_۲}{۲}\right) \leq \frac{f(x_۱) + f(x_۲)}{۲} \tag{۷.آ}$$

$$f(x) \geq f(y) + f'(y)(x-y) \tag{۸.آ}$$

$$f(tx) = f(tx + (۱-t)۰) \leq tf(x) + (۱-t)f(۰) \leq tf(x), \forall t \in [۰, ۱]. \tag{۹.آ}$$

$$\begin{aligned} f(a) + f(b) &= f\left((a+b)\frac{a}{a+b}\right) + f\left((a+b)\frac{b}{a+b}\right) \\ &\leq \frac{a}{a+b}f(a+b) + \frac{b}{a+b}f(a+b) = f(a+b) \end{aligned} \tag{۱۰.آ}$$

همچنین چنانچه تابع $f : (a, b) \to \mathbb{R}$ مشتق‌پذیر باشد، می‌توان نشان داد این تابع محدب است اگر و تنها اگر f' صعودی باشد. چنانچه f دارای مشتق دوم باشد، این بدان معناست که تابع f محدب است اگر و تنها اگر $f'' \geq ۰$.

۱.۱.آ سری تیلور

تعریف آ.۱.۱ (limsup)**.** دنباله X_n را در نظر بگیرید. در این صورت

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

یا به طور معادل، می‌توان فرض کرد دنباله‌ی J_n به صورت زیر تعریف شده باشد.

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

دنباله J_n یک دنباله ناصعودی ($J_n \supseteq J_{n+1}$) است، زیرا J_{n+1} اجتماع مجموعه‌های کوچکتر از J_n است. بزرگترین کران پایین روی این دنباله از اجتماع X_n ها برابر است با

$$\begin{aligned}\limsup_{n\to\infty} X_n &= inf\{sup\{X_m : m \in \{n, n+1, n+2, \dots\}\} : n \in \{1, 2, \dots\}\} \\ &= \cap_{n=1}^{\infty}(\cup_{m=n}^{\infty} X_m)\end{aligned}$$

تعریف آ.۲.۱. فرض کنید X_n و Y_n دو دنباله از متغیرهای تصادفی باشند. اگر $\limsup_{n\to\infty}|\frac{X_n}{Y_n}| < \infty$ به عبارت دیگر اگر $n \to \infty$، عبارت $|\frac{X_n}{Y_n}|$ کران‌دار باشد می‌نویسیم $X_n = O(Y_n)$.

تعریف آ.۳.۱. فرض کنید X_n و Y_n دو دنباله از متغیرهای تصادفی باشند. اگر

$$\lim_{n\to\infty}\left|\frac{X_n}{Y_n}\right| = 0$$

می‌نویسیم $X_n = o(Y_n)$.

در ریاضیات علامت O بزرگ رفتار حدی یک تابع را وقتی آرگومان‌های آن به یک عدد خاص یا به بی نهایت میل می‌کند را بیان می‌کند علامت O بزرگ به کاربر این اجازه را 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 را برحسب n در نظر بگیرید

$$T(n) = 4n^2 - 5n + 7$$

اگر تعداد ورودی این مسأله به بی‌نهایت میل کند اندازه جمله n^2 بسیار نزدیک تر از دیگر جمله‌ها خواهد بود. در این صورت گفته می‌شود $T(n) = O(n^2)$ یا $T(n) \in O(n^2)$.
در ریاضیات معمولا برای نشان دادن این که یک سری هندسی متناهی تا چه اندازه به تابع مورد نظر نزدیک است خصوصا در سری تیلور ناقص از این علامت استفاده می‌شود.

تعریف آ.۴.۱. فرض کنید f یک تابع حقیقی‌ـ‌مقدار روی R باشد و همچنین $x \in R$. اگر برای بعضی δ ها تابع f دارای مشتق‌های پیوسته تا مرتبه‌ی p در بازه $(x-\delta, x+\delta)$ باشد، آن‌گاه برای هر دنباله‌ی α_n همگرا به صفر، داریم

$$\begin{aligned}f(x+\alpha_n) &= \sum_{j=1}^{p}\left(\frac{\alpha_n^j}{j!}\right)f^{(j)}(x) + o(\alpha_n^p) \\ &= \sum_{j=1}^{p}\left(\frac{\alpha_n^j}{j!}\right)f^{(j)}(x) + O(\alpha_n^{p+1})\end{aligned}$$

۲.۱.آ روش اعتبارسنجی متقابل

بدیهی است که روش های جریمه‌شده، به پارامتر تنظیم‌کننده λ وابسته هستند پس انتخاب این پارامتر برای به کارگیری تابع جریمه از اهمیت زیادی برخوردار است. روش های زیادی برای یافتن آن وجود دارد اما روشی که بیشتر مورد استفاده قرار می‌گیرد روش اعتبارسنجی متقابل[۱] (CV) است . این روش برای تخمین بهترین مقدار λ به کار می‌رود. در این روش علاقه‌مندیم پارامتر λ را طوری پیدا کنیم که میانگین توان دوم خطای پیشگویی[۲] (MPE) کمترین مقدار را داشته باشد. لازم به ذکر است در مدل (۳.۱) معیار MPE به صورت زیر محاسبه می‌شود

$$MPE = E(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top (\boldsymbol{y} - \hat{\boldsymbol{y}})$$

که در آن $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. کار اعتبارسنجی متقابل به صورت زیر است. ابتدا داده‌ها را به k زیرگروه مساوی تقسیم می‌کنیم. یک گروه به عنوان برازش $\hat{\boldsymbol{\beta}}$ و $k-۱$ گروه به عنوان ارزیابی $\hat{y} = \hat{\boldsymbol{\beta}}x$ در نظر گرفته می‌شود که خطای مقدار برازش شده مشاهدات محاسبه می‌شود. این روند روی هر قسمت تکرار می‌شود و در آخر میانگین این خطاها محاسبه می‌شود. معمولا k را ۵، ۱۰و n در نظر می‌گیرند. در این حالت معیار CV به صورت زیر تعریف می‌شود

$$CV(\lambda) = \frac{۱}{n}\sum_{n}^{i=۱}(y_i - \hat{y_\lambda}^{-i})^۲$$

که در آن $\hat{Y_\lambda}^{-i}$ بیانگر مقدار برازش شده پس از کنار گذاشتن i امین گروه مشاهدات است. در این صورت λ مورد نظر متناظر با λیی خواهد بود که این معیار را مینیمم کند.

محاسبه $CV(\lambda)$ بسیار وقت گیر است. کروان و واهبا (۱۹۷۹) معیار اعتبارسنجی تعمیم‌یافته[۳] (GCV) را به عنوان جایگزینی برای $CV(\lambda)$ به صورت زیر معرفی کردند

$$GCV(\lambda) = \frac{(\boldsymbol{y} - \hat{\boldsymbol{Y}}_\lambda)^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\lambda)}{n(۱ - \frac{tr(A(\lambda))}{n})^۲}$$

که در آن، $\hat{Y}_\lambda$ بردار مقادیر برازش‌شده تحت ماتریس تصویر $A(\lambda)$ است. انتخاب پارامتر جریمه بخش مهمی از برازش مدل در رگرسیون جریمه‌شده است. روش‌های بسیاری در این مورد وجود دارد اما مهمترین روش‌های کاربردی بر اساس این است که برآوردگر

[۱] Cross-validation
[۲] Mean prediction error
[۳] Generalized cross-validation

جریمه‌شده چه‌قدر خوب می‌تواند مقدار واقعی متغیر پاسخ را پیش‌گویی کند. منصفانه نیست از داده‌ها دوبار استفاده شود، یک بار برای برازش مدل و دوباره از همان داده‌ها برای برآورد دقت پیشگویی. این کار، منجر به بیش‌برازشی می‌شود.

ایده‌ای که به‌ذهن می‌رسد، تقسیم‌بندی داده‌ها به دو مجموعه است. از یک مجموعه (مجموعه آموزش) برای برازش برآوردگر جریمه‌شده (مثلا β^*) استفاده شده و از مجموعه دیگر (مجموعه آزمون) برای محاسبه اینکه کمیت $\boldsymbol{X}\beta^*$ چه‌قدر خوب توانسته است مشاهدات مجموعه دیگر را پیش‌گویی کند. اما به‌ندرت تعداد داده‌ها آن‌قدر زیاد است که بتوان آن‌ها را به دو قسمت تقسیم کرد به‌طوری که هر یک از مجموعه‌ها به‌تنهایی برای انتخاب پارامتر جریمه کافی باشند. از این رو، روش اعتبارسنجی متقابل پیشنهاد می‌شود.

در اعتبارسنجی متقابل، داده‌ها به‌طور تصادفی به K زیرمجموعه تقسیم می‌شود. یک زیرمجموعه کنار گذاشته می‌شود که نام مجموعه آزمون را به خود می‌گیرد، $K-1$ مجموعه باقیمانده، مجموعه آموزشی نامیده و برای برازش مدل استفاده می‌گردد. سپس مدل برازش شده به‌منظور پیش‌گویی متغیرهای پاسخ در مجموعه داده آزمون به‌کار برده می‌شود. سرانجام، خطاهای پیش‌گویی با میانگین توان دوم انحراف مقادیر مشاهده شده از مقادیر پیش‌گویی در مجموعه داده آزمون به‌دست می‌آیند. معمولا $K = ۵, ۱۰, n$ انتخاب می‌شود.

آ.۲ قضیه لایب نیتز

مشتق انتگرال: هرگاه بخواهیم از تابعی شامل انتگرال، مشتق بگیریم باید از قاعده زیر استفاده کنیم که مستقیماً فرمول محاسبه مشتق انتگرال را به ما می‌دهد. مشتق انتگرال $\int_{m(x)}^{n(x)} h(t)dt$ برابر است با:

$$\left(\int_{m(x)}^{n(x)} h(t)dt\right)' = m'(x).h(m(x)) - n'(x).h(n(x)) \qquad (آ.۱۱)$$

یعنی مشتق تابعی که شامل انتگرالی با یک تابع تک‌متغیره مانند $h(t)$ و کران‌هایی بر حسب x مانند $m(x)$ (کران بالا) و $n(x)$ (کران پایین) باشد برابر است با مشتق کران بالا ضربدر تابع داخل انتگرال که کران بالا به جای متغیر آن جایگزین شده است منهای مشتق کران پایین ضربدر تابع داخل انتگرال که کران پایین به جای متغیر آن جایگزین شده باشد.

مثال آ.۱.۲. مشتق انتگرال

$$f(x) = \int_{Ln(x)}^{x^{۳}+۲x} \frac{۲t}{۳-t} dt \qquad (۱۲.آ)$$

را بیابید.

حل.

$$\begin{aligned} f(x) &= \int_{Ln(x)}^{x^{۳}+۲x} \frac{۲t}{۳-t} dt \\ \Rightarrow f'(x) &= (۳x^{۲}+۲).\frac{۲(x^{۳}+۲x)}{۳-(x^{۳}+۲x)} - \frac{۱}{x}.\frac{۲Ln(x)}{۳-Ln(x)} \end{aligned}$$

□

قضیه آ.۱.۲ (قضیه انتگرال گیری جزءبه جزء)**.** فرض کنید u و v توابعی دیفرانسیل‌پذیر باشد در این صورت دیفرانسیل تابع uv به صورت زیر محاسبه می‌شود

$$d(uv) = udv + vdu$$

که می‌توان آن را به صورت زیر نوشت

$$udv = d(uv) - vdu$$

پس از انتگرال‌گیری از آن داریم

$$\int udv = uv - \int vdu \qquad (۱۳.آ)$$

در فرمول انتگرال‌گیری جزء به جزء، انتگرال $\int udv$ برحسب انتگرال $\int vdu$ بیان می‌شود. با انتخاب مناسب u و v ممکن است محاسبه انتگرال دوم ساده‌تر از محاسبه انتگرال اول باشد. اهمیت این فرمول در همین است. وقتی با انتگرالی رو به رو می‌شویم که نمی‌توانیم آن را حل کنیم، باید آن را به انتگرال دیگری تبدیل کنیم تا بتوانیم آن را حل کنیم.

۱.۲.آ معیار AIC و BIC

معیار آکائیک (AIC) به صورت زیر تعریف می‌شود

$$AIC = -۲l + ۲p$$

که در آن، l لگاریتم تابع درست نمایی و p تعداد ضریب رگرسیونی موجود در مدل می‌باشد. هر مدل زیرمجموعه با AIC کمتر، بر سایر مدل‌های زیر مجموعه برتری دارد. پس از این معیار، معیارهای دیگری پدید آمده اند که معروف‌ترین آنها BIC است و به صورت زیر تعریف می‌شود

$$BIC = -۲l + p\ln(n)$$

که در آن، n حجم نمونه است.

پیوست ب

گزیده‌ای از برنامه‌های رایانه‌ای

نمودار ۱.۲

```
library(philentropy)
library(latex2exp)
X=seq(-2,2,0.05)
s = seq(0.0001,3,length.out = 1000)


F1 = function(x,s)     sqrt((x^2)+s)
F2 = function(x,s)   sqrt((x^2)+(s^2))
F3 = function(x,s)    (x^2)/sqrt((x^2)+(s^2))
F4 = function(x,s) log(2+exp(-x/s)+exp(x/s))*s
F5 = function(x,s)   x*(2*pnorm(x/s,0,1/sqrt(2),lower.tail = TRUE)-1)

D1 <- D2 <- D3 <- D4 <- D5 <- numeric(length(s))
for(i in 1:length(D1)){
```

```
D1[i] = distance(rbind(abs(X),F1(X,s[i])), method = "euclidean")
D2[i] = distance(rbind(abs(X),F2(X,s[i])), method = "euclidean")
D3[i] = distance(rbind(abs(X),F3(X,s[i])), method = "euclidean")
D4[i] = distance(rbind(abs(X),F4(X,s[i])), method = "euclidean")
D5[i] = distance(rbind(abs(X),F5(X,s[i])), method = "euclidean")
}
plot(s,D1,type = "l",ylim = c(0,28),lwd = 1.5 , col = "black",ylab = "distance")
lines(s,D2,type = "l",lwd = 1.5 , col = "green")
lines(s,D3,type = "l",lwd = 1.5 , col = "blue")
lines(s,D4,type = "l",lwd = 1.5 , col = "purple")
lines(s,D5,type = "l",lwd = 2 , col = "red")

expression.legend = c(TeX("$\\sqrt{x^2+2}$"),TeX("$\\sqrt{x^2+s^2}$"),
TeX("$x^2/\\sqrt{x^2+s^2}$"),TeX("$\\log(2+exp(-x/s)+exp(x/s))s)$"),
TeX("$x(2 \\Phi(x/s,0,1/\\sqrt{2})-1)$"))
legend(0.05,28,expression.legend , col = c("black","green","blue","purple","red"),
lwd = c(1.5,1.5,1.5,1.5,2), lty = 1)

x = seq(-1,1,BY = 0.05)
x2= function(x) x^2
s1 = 0.01
f1 = function(x)   x*(2*pnorm(x/s1,0,1/sqrt(2))-1)
s2=2/sqrt(pi)
f2=function(x)     x*(2*pnorm(x/s2,0,1/sqrt(2))-1)

curve(f2,-1,1,col = "red",lwd = 2,xlab = "x" ,ylab = "p(x,s)",ylim = c(0,1.1))
curve(f1, -1,1,col = "orange" , lwd = 2, add= TRUE)
curve(x2, -1,1,col = "blue" , lwd = 1.5,add = TRUE)
curve(abs(x),-1,1,col="black",lwd=2,add=TRUE , lty = 3)

expression.legend = c(TeX("$x(2 \\Phi(x/s,0,1/\\sqrt{2})-1), s = 2/\\sqrt{\\pi}$"),
TeX("$x(2 \\Phi(x/s,0,1/\\sqrt{2})-1), s = 0.01$"),TeX("$x^2$"),TeX("$|x|$"))
```

```
legend(-0.5,1,col=c("red","green","blue","black"), expression.legend,
lty = c(1,1,1,3),lwd=c(2,2,1.5,2),text.width = 0.7)
```

نمودار ۲.۲

```
library(latex2exp)
thresholding.function <- function(beta,s = 0.01){
.5*(2*pnorm(beta/s,0,1/sqrt(2))-1+2*(beta/s)*dnorm(beta/s,0,1/sqrt(2)))
+ beta }
f.inv <- inverse(thresholding.function, -1,1)

y.vec = seq(-1.5,1.5,0.01)

D = numeric(length(y.vec))
for(i in 1:length(y.vec))
D[i] = f.inv(y.vec[i])
plot(y.vec,D, type = "l",lwd = 2 , xlab = TeX("$x_i$") ,
ylab = TeX("$\\hat{\\beta}_i"))
title(TeX("$ s = 0.01 (lasso) $"))
abline(a = 0 , b = 1 , lty = 3)
abline(h = 0, lty = 2)
abline(v = -0.5, lty = 2)
abline(v = 0.5, lty = 2)

thresholding.function <- function(beta,s = 0.5){
.5*(2*pnorm(beta/s,0,1/sqrt(2))-1+2*(beta/s)*dnorm(beta/s,0,1/sqrt(2))) + beta
}
f.inv <- inverse(thresholding.function, -1,1)

y.vec = seq(-1.5,1.5,0.01)

D = numeric(length(y.vec))
for(i in 1:length(y.vec))
```

```
D[i] = f.inv(y.vec[i])
plot(y.vec,D, type = "l",lwd = 2 , xlab = TeX("$x_i$") ,
 ylab = TeX("$\\hat{\\beta}_i"))
title(TeX("$ s = 0.5 $"))
abline(a = 0 , b = 1 , lty = 3)
abline(h = 0, lty = 2)
abline(v = -0.5, lty = 2)
abline(v = 0.5, lty = 2)

thresholding.function <- function(beta,s = 1){
.5*(2*pnorm(beta/s,0,1/sqrt(2))-1+2*(beta/s)*dnorm(beta/s,0,1/sqrt(2)))
+ beta }
f.inv <- inverse(thresholding.function, -1,1)

y.vec = seq(-1.5,1.5,0.01)

D = numeric(length(y.vec))
for(i in 1:length(y.vec))
D[i] = f.inv(y.vec[i])
plot(y.vec,D, type = "l",lwd = 2 , xlab = TeX("$x_i$") ,
 ylab = TeX("$\\hat{\\beta}_i"))
title(TeX("$ s = 1 (ridge) $"))
abline(a = 0 , b = 1 , lty = 3)
abline(h = 0, lty = 2)
abline(v = -0.5, lty = 2)
abline(v = 0.5, lty = 2)

thresholding.function <- function(beta,s = 20){
.5*(2*pnorm(beta/s,0,1/sqrt(2))-1+2*(beta/s)*dnorm(beta/s,0,1/sqrt(2))) + beta
}
f.inv <- inverse(thresholding.function, -2,2)
```

```
y.vec = seq(-1.5,1.5,0.01)

D = numeric(length(y.vec))
for(i in 1:length(y.vec))
D[i] = f.inv(y.vec[i])
plot(y.vec,D, type = "l",lwd = 2 , xlab = TeX("$x_i$") ,
 ylab = TeX("$\\hat{\\beta}_i"))
title(TeX("$ s = 20  (OLS) $"))
abline(a = 0 , b = 1 , lty = 3)
abline(h = 0, lty = 2)
abline(v = -0.5, lty = 2)
abline(v = 0.5, lty = 2)
```

نمودار ۳.۲

```
library(MASS)
library(DLASSO)
library(ncvreg)
library(glmnet)
library(msgps)
library(rms)
library(lasso2)

#Senario 1
sigm2 = 3
s = sqrt(sigm2)
p = 8
beta = matrix(c(3,1.5,0,0,2,0,0,0),nr = 8, nc = 1)
Sigma = matrix(0,nr = p, nc = p)

for(i in 1:p){
for(j in 1:p){
Sigma[i,j] = 0.5^(abs(i-j))
```

```
}
}

n = 240  #number of observations
mu = rep(0,p)
X = mvrnorm(n  , mu = mu, Sigma = Sigma)
X.train = X[1:40,]
X.test = X[41:240,]

N.Sim = 50
DLASSO.mat <- OLS.mat <- ridge.mat <- lasso.mat <-enet.mat <-
SCAD.mat <- matrix(0,nr = p, nc = N.Sim)
for(i.sim in 1:N.Sim){
e = rnorm(n,mean = 0, sd = 1) #error
e.train = e[1:40]
e.test = e[41:240]
y.train = X.train%*%beta + s*e.train
y.train <- c(y.train)
y.test = X.test%*%beta + s*e.test
dLasso = dlasso(x=X.train,y=y.train,adp = FALSE , s = 0.01)
DLASSO.mat[,i.sim] = coef(dLasso)[,4]
ridge=cv.glmnet(x=X.train,y=y.train,alpha = 0)
ridge.mat[,i.sim] = coef(ridge,s = ridge$lambda.min)[-1]
lasso=fit <- msgps(X.train,y.train)
lasso.mat[,i.sim]=coef(lasso)[-1,4]
enet=msgps(X.train,y.train,penalty="enet",alpha=0.5)
enet.mat[,i.sim]=coef(enet)[-1,4]
scad = cv.ncvreg(X.train,y.train,penalty = "SCAD")
SCAD.mat[,i.sim]<-ncvreg(X.train,y.train,penalty = "SCAD",
lambda = scad$lambda.min)$beta[-1]
ols = lm(y.train ~ X.train -1)
OLS.mat[,i.sim] = coef(ols)
```

```
}

y.ols <- y.scad <- y.dlasso <- y.enet <- y.lasso <-
 y.ridge <- matrix(0, nr = 200, nc = N.Sim)
for(i.test in 1:N.Sim){
y.ols[,i.test] <- X.test %*% OLS.mat[,i.test]
y.scad[,i.test] <- X.test %*% SCAD.mat[,i.test]
y.dlasso[,i.test] <- X.test %*% DLASSO.mat[,i.test]
y.enet[,i.test] <- X.test %*% enet.mat[,i.test]
y.lasso[,i.test] <- X.test %*% lasso.mat[,i.test]
y.ridge[,i.test] <- X.test %*% ridge.mat[,i.test]
}

MSE <- matrix(0, nr = N.Sim, nc = 6)
for(i.mse in 1:N.Sim){
MSE[i.mse,1] <- sum((y.ols[,i.mse]-y.test)^2)/200  #OLS
MSE[i.mse,2] <- sum((y.ridge[,i.mse]-y.test)^2)/200  #ridge
MSE[i.mse,3] <- sum((y.dlasso[,i.mse]-y.test)^2)/200  #DLASSO
MSE[i.mse,4] <- sum((y.lasso[,i.mse]-y.test)^2)/200  #lasso
MSE[i.mse,5] <- sum((y.enet[,i.mse]-y.test)^2)/200  #enet
MSE[i.mse,6] <- sum((y.scad[,i.mse]-y.test)^2)/200  #scad
}
colnames(MSE)<- c("OLS","Ridge","DLASSO","LASSO","ENET","SCAD")
boxplot(MSE,ylim = c(2,6))

MSE.beta <- matrix(0, nr = N.Sim, nc = 6)
S.X <- cov(X)
for(i.mse.beta in 1:N.Sim){
MSE.beta[i.mse.beta,1] <- t(OLS.mat[,i.mse.beta]-beta)%*%S.X%*%
(OLS.mat[,i.mse.beta]-beta)#OLS
MSE.beta[i.mse.beta,2] <- t(ridge.mat[,i.mse.beta]-beta)%*%S.X%*%
(ridge.mat[,i.mse.beta]-beta)  #ridge
```

```
MSE.beta[i.mse.beta,3] <- t(DLASSO.mat[,i.mse.beta]-beta)%*%S.X%*%
(DLASSO.mat[,i.mse.beta]-beta)  #DLASSO
MSE.beta[i.mse.beta,4] <- t(lasso.mat[,i.mse.beta]-beta)%*%S.X%*%
(lasso.mat[,i.mse.beta]-beta)  #lasso
MSE.beta[i.mse.beta,5] <- t(enet.mat[,i.mse.beta]-beta)%*%S.X%*%
(enet.mat[,i.mse.beta]-beta)  #enet
MSE.beta[i.mse.beta,6] <- t(SCAD.mat[,i.mse.beta]-beta)%*%S.X%*%
(SCAD.mat[,i.mse.beta]-beta)  #scad
}
colnames(MSE.beta)<- c("OLS","Ridge","DLASSO","LASSO","ENET","SCAD")
boxplot(MSE.beta,ylim = c(0,3))

#Senario 2
sigm2 = 3
s = sqrt(sigm2)
p = 8
beta2=rep(0.5,8)
Sigma = matrix(0,nr = p, nc = p)

for(i in 1:p){
for(j in 1:p){
Sigma[i,j] = 0.5^(abs(i-j))
}
}

n = 240  #number of observations
mu = rep(0,p)
X = mvrnorm(n  , mu = mu, Sigma = Sigma)
X.train = X[1:40,]
X.test = X[41:240,]

N.Sim = 50
```

```
DLASSO.mat <- OLS.mat <- ridge.mat <- lasso.mat <-enet.mat <-SCAD.mat <-
 matrix(0,nr = p, nc = N.Sim)
for(i.sim in 1:N.Sim){
e = rnorm(n,mean = 0, sd = 1) #error
e.train = e[1:40]
e.test = e[41:240]
y.train = X.train%*%beta2 + s*e.train
y.train <- c(y.train)
y.test = X.test%*%beta2 + s*e.test
dLasso = dlasso(x=X.train,y=y.train,adp = FALSE , s = 0.01)
DLASSO.mat[,i.sim] = coef(dLasso)[,4]
ridge=cv.glmnet(x=X.train,y=y.train,alpha = 0)
ridge.mat[,i.sim] = coef(ridge,s = ridge$lambda.min)[-1]
lasso=fit <- msgps(X.train,y.train)
lasso.mat[,i.sim]=coef(lasso)[-1,4]
enet=msgps(X.train,y.train,penalty="enet",alpha=0.5)
enet.mat[,i.sim]=coef(enet)[-1,4]
scad = cv.ncvreg(X.train,y.train,penalty = "SCAD")
SCAD.mat[,i.sim]<-ncvreg(X.train,y.train,penalty = "SCAD",
lambda = scad$lambda.min)$beta[-1]
ols = lm(y.train ~ X.train -1)
OLS.mat[,i.sim] = coef(ols)
}

y.ols <- y.scad <- y.dlasso <- y.enet <- y.lasso <- y.ridge <-
 matrix(0, nr = 200, nc = N.Sim)
for(i.test in 1:N.Sim){
y.ols[,i.test] <- X.test %*% OLS.mat[,i.test]
y.scad[,i.test] <- X.test %*% SCAD.mat[,i.test]
y.dlasso[,i.test] <- X.test %*% DLASSO.mat[,i.test]
y.enet[,i.test] <- X.test %*% enet.mat[,i.test]
y.lasso[,i.test] <- X.test %*% lasso.mat[,i.test]
```

```
y.ridge[,i.test] <- X.test %*% ridge.mat[,i.test]
}

MSE <- matrix(0, nr = N.Sim, nc = 6)
for(i.mse in 1:N.Sim){
MSE[i.mse,1] <- sum((y.ols[,i.mse]-y.test)^2)/200  #OLS
MSE[i.mse,2] <- sum((y.ridge[,i.mse]-y.test)^2)/200  #ridge
MSE[i.mse,3] <- sum((y.dlasso[,i.mse]-y.test)^2)/200  #DLASSO
MSE[i.mse,4] <- sum((y.lasso[,i.mse]-y.test)^2)/200  #lasso
MSE[i.mse,5] <- sum((y.enet[,i.mse]-y.test)^2)/200  #enet
MSE[i.mse,6] <- sum((y.scad[,i.mse]-y.test)^2)/200  #scad
}
colnames(MSE)<- c("OLS","Ridge","DLASSO","LASSO","ENET","SCAD")
boxplot(MSE,ylim = c(2,6))

MSE.beta <- matrix(0, nr = N.Sim, nc = 6)
S.X <- cov(X)
for(i.mse.beta in 1:N.Sim){
MSE.beta[i.mse.beta,1] <- t(OLS.mat[,i.mse.beta]-beta2)%*%S.X%*%
(OLS.mat[,i.mse.beta]-beta2)  #OLS
MSE.beta[i.mse.beta,2] <- t(ridge.mat[,i.mse.beta]-beta2)%*%S.X%*%
(ridge.mat[,i.mse.beta]-beta2)  #ridge
MSE.beta[i.mse.beta,3] <- t(DLASSO.mat[,i.mse.beta]-beta2)%*%S.X%*%
(DLASSO.mat[,i.mse.beta]-beta2)  #DLASSO
MSE.beta[i.mse.beta,4] <- t(lasso.mat[,i.mse.beta]-beta2)%*%S.X%*%
(lasso.mat[,i.mse.beta]-beta2)  #lasso
MSE.beta[i.mse.beta,5] <- t(enet.mat[,i.mse.beta]-beta2)%*%S.X%*%
(enet.mat[,i.mse.beta]-beta2)  #enet
MSE.beta[i.mse.beta,6] <- t(SCAD.mat[,i.mse.beta]-beta2)%*%S.X%*%
(SCAD.mat[,i.mse.beta]-beta2)  #scad
}
colnames(MSE.beta)<- c("OLS","Ridge","DLASSO","LASSO","ENET","SCAD")
```

```
boxplot(MSE.beta,ylim = c(0,4))
########################################
#Senario 3
sigma3=15
s=sqrt(sigma3)
p=15
beta1=c(1,2,3,4,5)
beta2=rep(0.5,5)
beta3=rep(0,5)
beta=matrix(c(beta1,beta2,beta3),nr=p,nc=1)
Sigma = matrix(0,nr = p, nc = p)

for(i in 1:p){
for(j in 1:p){
Sigma[i,j] = 0.5^(abs(i-j))
}
}

n = 240  #number of observations
mu = rep(0,p)
X = mvrnorm(n  , mu = mu, Sigma = Sigma)
X.train = X[1:40,]
X.test = X[41:240,]

N.Sim = 50
DLASSO.mat <- OLS.mat <- ridge.mat <- lasso.mat <-enet.mat <-SCAD.mat <-
matrix(0,nr = p, nc = N.Sim)
for(i.sim in 1:N.Sim){
e = rnorm(n,mean = 0, sd = 1) #error
e.train = e[1:40]
e.test = e[41:240]
y.train = X.train%*%beta + s*e.train
```

```
y.train <- c(y.train)
y.test = X.test%*%beta + s*e.test
dLasso = dlasso(x=X.train,y=y.train,adp = FALSE , s = 0.01)
DLASSO.mat[,i.sim] = coef(dLasso)[,4]
ridge=cv.glmnet(x=X.train,y=y.train,alpha = 0)
ridge.mat[,i.sim] = coef(ridge,s = ridge$lambda.min)[-1]
lasso=fit <- msgps(X.train,y.train)
lasso.mat[,i.sim]=coef(lasso)[-1,4]
enet=msgps(X.train,y.train,penalty="enet",alpha=0.5)
enet.mat[,i.sim]=coef(enet)[-1,4]
scad = cv.ncvreg(X.train,y.train,penalty = "SCAD")
SCAD.mat[,i.sim]<-ncvreg(X.train,y.train,penalty = "SCAD",
lambda = scad$lambda.min)$beta[-1]
ols = lm(y.train ~ X.train -1)
OLS.mat[,i.sim] = coef(ols)
}

y.ols <- y.scad <- y.dlasso <- y.enet <- y.lasso <-
y.ridge <- matrix(0, nr = 200, nc = N.Sim)
for(i.test in 1:N.Sim){
y.ols[,i.test] <- X.test %*% OLS.mat[,i.test]
y.scad[,i.test] <- X.test %*% SCAD.mat[,i.test]
y.dlasso[,i.test] <- X.test %*% DLASSO.mat[,i.test]
y.enet[,i.test] <- X.test %*% enet.mat[,i.test]
y.lasso[,i.test] <- X.test %*% lasso.mat[,i.test]
y.ridge[,i.test] <- X.test %*% ridge.mat[,i.test]
}

MSE <- matrix(0, nr = N.Sim, nc = 6)
for(i.mse in 1:N.Sim){
MSE[i.mse,1] <- sum((y.ols[,i.mse]-y.test)^2)/200  #OLS
MSE[i.mse,2] <- sum((y.ridge[,i.mse]-y.test)^2)/200  #ridge
```

```
MSE[i.mse,3] <- sum((y.dlasso[,i.mse]-y.test)^2)/200  #DLASSO
MSE[i.mse,4] <- sum((y.lasso[,i.mse]-y.test)^2)/200  #lasso
MSE[i.mse,5] <- sum((y.enet[,i.mse]-y.test)^2)/200  #enet
MSE[i.mse,6] <- sum((y.scad[,i.mse]-y.test)^2)/200  #scad
}
colnames(MSE)<- c("OLS","Ridge","DLASSO","LASSO","ENET","SCAD")
boxplot(MSE,ylim = c(15,40))

MSE.beta <- matrix(0, nr = N.Sim, nc = 6)
S.X <- cov(X)
for(i.mse.beta in 1:N.Sim){
MSE.beta[i.mse.beta,1] <- t(OLS.mat[,i.mse.beta]-beta)%*%S.X%*%
(OLS.mat[,i.mse.beta]-beta)  #OLS
MSE.beta[i.mse.beta,2] <- t(ridge.mat[,i.mse.beta]-beta)%*%S.X%*%
(ridge.mat[,i.mse.beta]-beta)  #ridge
MSE.beta[i.mse.beta,3] <- t(DLASSO.mat[,i.mse.beta]-beta)%*%S.X%*%
(DLASSO.mat[,i.mse.beta]-beta)  #DLASSO
MSE.beta[i.mse.beta,4] <- t(lasso.mat[,i.mse.beta]-beta)%*%S.X%*%
(lasso.mat[,i.mse.beta]-beta)  #lasso
MSE.beta[i.mse.beta,5] <- t(enet.mat[,i.mse.beta]-beta)%*%S.X%*%
(enet.mat[,i.mse.beta]-beta)  #enet
MSE.beta[i.mse.beta,6] <- t(SCAD.mat[,i.mse.beta]-beta)%*%S.X%*%
(SCAD.mat[,i.mse.beta]-beta)  #scad
}
colnames(MSE.beta)<- c("OLS","Ridge","DLASSO","LASSO","ENET","SCAD")
boxplot(MSE.beta,ylim = c(0,20))
########################################
#Senario 4
sigm2 = 3
s = sqrt(sigm2)
p = 300
u=rep(c(3,1.5,0,0,2,0),c(1,1,1,1,1,295))
```

```
beta = matrix(u,nr = 300, nc = 1)
Sigma = matrix(0,nr = p, nc = p)

for(i in 1:p){
for(j in 1:p){
Sigma[i,j] = 0.5^(abs(i-j))
}
}

n = 240  #number of observations
mu = rep(0,p)
X = mvrnorm(n  , mu = mu, Sigma = Sigma)
X.train = X[1:40,]
X.test = X[41:240,]

N.Sim = 50
DLASSO.mat <- OLS.mat <- ridge.mat <- lasso.mat <-enet.mat <-
SCAD.mat <- matrix(0,nr = p, nc = N.Sim)
for(i.sim in 1:N.Sim){
e = rnorm(n,mean = 0, sd = 1) #error
e.train = e[1:40]
e.test = e[41:240]
y.train = X.train%*%beta + s*e.train
y.train <- c(y.train)
y.test = X.test%*%beta + s*e.test
dLasso = dlasso(x=X.train,y=y.train,adp = FALSE , s = 0.01)
DLASSO.mat[,i.sim] = coef(dLasso)[,4]
ridge=cv.glmnet(x=X.train,y=y.train,alpha = 0)
ridge.mat[,i.sim] = coef(ridge,s = ridge$lambda.min)[-1]
lasso=fit <- msgps(X.train,y.train)
lasso.mat[,i.sim]=coef(lasso)[-1,4]
enet=msgps(X.train,y.train,penalty="enet",alpha=0.5)
```

```
enet.mat[,i.sim]=coef(enet)[-1,4]
scad = cv.ncvreg(X.train,y.train,penalty = "SCAD")
SCAD.mat[,i.sim]<-ncvreg(X.train,y.train,penalty = "SCAD",
lambda = scad$lambda.min)$beta[-1]
ols = lm(y.train ~ X.train -1)
OLS.mat[,i.sim] = coef(ols)
}

y.ols <- y.scad <- y.dlasso <- y.enet <- y.lasso <-
y.ridge <- matrix(0, nr = 200, nc = N.Sim)
for(i.test in 1:N.Sim){



y.ols[,i.test] <- X.test %*% OLS.mat[,i.test]
y.scad[,i.test] <- X.test %*% SCAD.mat[,i.test]
y.dlasso[,i.test] <- X.test %*% DLASSO.mat[,i.test]
y.enet[,i.test] <- X.test %*% enet.mat[,i.test]
y.lasso[,i.test] <- X.test %*% lasso.mat[,i.test]
y.ridge[,i.test] <- X.test %*% ridge.mat[,i.test]
}

MSE <- matrix(0, nr = N.Sim, nc = 6)
for(i.mse in 1:N.Sim){
MSE[i.mse,1] <- sum((y.ols[,i.mse]-y.test)^2)/200  #OLS
MSE[i.mse,2] <- sum((y.ridge[,i.mse]-y.test)^2)/200  #ridge
MSE[i.mse,3] <- sum((y.dlasso[,i.mse]-y.test)^2)/200  #DLASSO
MSE[i.mse,4] <- sum((y.lasso[,i.mse]-y.test)^2)/200  #lasso
MSE[i.mse,5] <- sum((y.enet[,i.mse]-y.test)^2)/200  #enet
MSE[i.mse,6] <- sum((y.scad[,i.mse]-y.test)^2)/200  #scad
}
colnames(MSE)<- c("OLS","Ridge","DLASSO","LASSO","ENET","SCAD")
boxplot(MSE,ylim = c(2,22))
```

```
MSE.beta <- matrix(0, nr = N.Sim, nc = 6)
S.X <- cov(X)
for(i.mse.beta in 1:N.Sim){
MSE.beta[i.mse.beta,1] <- t(OLS.mat[,i.mse.beta]-beta)%*%S.X%*%
(OLS.mat[,i.mse.beta]-beta)#OLS
MSE.beta[i.mse.beta,2] <- t(ridge.mat[,i.mse.beta]-beta)%*%S.X%*%
(ridge.mat[,i.mse.beta]-beta)  #ridge
MSE.beta[i.mse.beta,3] <- t(DLASSO.mat[,i.mse.beta]-beta)%*%S.X%*%
(DLASSO.mat[,i.mse.beta]-beta)  #DLASSO
MSE.beta[i.mse.beta,4] <- t(lasso.mat[,i.mse.beta]-beta)%*%S.X%*%
(lasso.mat[,i.mse.beta]-beta)  #lasso
MSE.beta[i.mse.beta,5] <- t(enet.mat[,i.mse.beta]-beta)%*%S.X%*%
(enet.mat[,i.mse.beta]-beta)  #enet
MSE.beta[i.mse.beta,6] <- t(SCAD.mat[,i.mse.beta]-beta)%*%S.X%*%
(SCAD.mat[,i.mse.beta]-beta)  #scad
}
colnames(MSE.beta)<- c("OLS","Ridge","DLASSO","LASSO","ENET","SCAD")
boxplot(MSE.beta,ylim = c(0,20))
```

شکل ۱.۳

```
n = 2000
t = numeric(n)
for(i in 1:n){
t[i] = (i - 0.5)/n
}
f = function(x)
sqrt(x*(1-x))*sin((2.1*pi)/(x+0.05))
f <- Vectorize(f)
f.t = matrix(f(t),nc = 1)

plot(t,f.t,type = "l",xlab = "t", ylab = "f")
```

شکل ۲.۳

```
library(mvtnorm)
library(DLASSO)
library(KernSmooth)

n = 100
p = 5

N.sim = 1000

beta = matrix(c(1.5,2,3,-5,4),nc = 1)
mu = c(-2,1,2,3,2)
Sigma = matrix(c(0.81,0.4,0.3,-0.2,-0.1,0.4,2.25,0.4,0.3,-0.2,
0.3,0.4,1,0.4,0.3,-0.2,0.3,0.4,0.64,0.4,
-0.1,-0.2,0.3,0.4,0.49), nr = 5 , nc = 5 , byrow = TRUE)
X = rmvnorm(n,mean = mu, sigma = Sigma)
t = numeric(n)
for(i in 1:n){
t[i] = (i - 0.5)/n }
f = function(x)
sqrt(x*(1-x))*sin((2.1*pi)/(x+0.05))
f <- Vectorize(f)
f.t = matrix(f(t),nc = 1)

V = matrix(0, nr = n, nc = n)
for(i in 1:n){
for(j in 1:n){
V[i,j] = (1/n)^(abs(i-j)+1) } }

sigma.e = 0.01
S.V = sigma.e * V
```

```
mu.e = rep(0,n)

d = c(0.8873,-0.3099,-0.2464,-0.1901,-0.1409)
m = length(d) -1
I.nm <- diag(1,n-m,n-m)
D = matrix(0,nr = (n-m), nc = n)
for(i in 1:(n-m)){
D[i,i:(i+m)] <- d  }

DLS.mat <- DDLASSO.mat <- matrix(0,nr = p,nc = N.sim)
e.DLS <- e.DDLASSO <- matrix(0,nr = n, nc = N.sim)
for(i.sim in 1:N.sim){
e = matrix(rmvnorm(1,mu.e,S.V),nc = 1)
y = X%*%beta + f.t + e

y.D = D %*% y
X.D = D %*% X

V.D = D %*% S.V %*% t(D)
C.D = t(X.D)%*%solve(V.D)%*%X.D
DLS.mat[,i.sim] = solve(C.D)%*%t(X.D)%*%solve(V.D)%*%y.D

dLasso = dlasso(x=X.D,y=y.D,adp = FALSE , s = 0.01)
DDLASSO.mat[,i.sim]  =  coef(dLasso)[,4]

e.DLS[,i.sim] <- y - X%*%DLS.mat[,i.sim]
e.DDLASSO[,i.sim] <- y - X%*%DDLASSO.mat[,i.sim]  }

R.DLS <- R.DDLASSO <- numeric(N.sim)
for(i.sim in 1:N.sim){
R.DLS[i.sim] <- t(DLS.mat[,i.sim]-beta)%*%(DLS.mat[,i.sim]-beta)
R.DDLASSO[i.sim] <- t(DDLASSO.mat[,i.sim]-beta)%*%(DDLASSO.mat[,i.sim]-beta)}
```

```
R<-cbind(R.DLS,R.DDLASSO)
colnames(R)<- c("DGLSE","DDLASSO")
boxplot(R)
```

شکل ۳.۳

```
library(mvtnorm)
library(DLASSO)
library(KernSmooth)
n = 2000
p = 5

beta = matrix(c(1.5,2,3,-5,4),nc = 1)
mu = c(-2,1,2,3,2)
Sigma = matrix(c(0.81,0.4,0.3,-0.2,-0.1,
0.4,2.25,0.4,0.3,-0.2,
0.3,0.4,1,0.4,0.3,
-0.2,0.3,0.4,0.64,0.4,
-0.1,-0.2,0.3,0.4,0.49),
nr = 5 , nc = 5 , byrow = TRUE)
X = rmvnorm(n,mean = mu, sigma = Sigma)
t = numeric(n)
for(i in 1:n){
t[i] = (i - 0.5)/n }
f = function(x)
sqrt(x*(1-x))*sin((2.1*pi)/(x+0.05))
f <- Vectorize(f)
f.t = matrix(f(t),nc = 1)

V = matrix(0, nr = n, nc = n)
for(i in 1:n){
```

```
for(j in 1:n){
V[i,j] = (1/n)^(abs(i-j)+1) } }

sigma.e = 0.01
S.V = sigma.e * V
mu.e = rep(0,n)

d = c(0.8873,-0.3099,-0.2464,-0.1901,-0.1409)
m = length(d) -1
I.nm <- diag(1,n-m,n-m)
D = matrix(0,nr = (n-m), nc = n)
for(i in 1:(n-m)){
D[i,i:(i+m)] <- d  }


e = matrix(rmvnorm(1,mu.e,S.V),nc = 1)
y = X%*%beta + f.t + e

y.D = D %*% y
X.D = D %*% X

V.D = D %*% S.V %*% t(D)
C.D = t(X.D)%*%solve(V.D)%*%X.D
DLS = solve(C.D)%*%t(X.D)%*%solve(V.D)%*%y.D

dLasso = dlasso(x=X.D,y=y.D,adp = FALSE , s = 0.01)
DDLASSO  =  coef(dLasso)[,4]

e.DLS <- y - X%*%DLS.mat[,i.sim]
e.DDLASSO <- y - X%*%DDLASSO.mat[,i.sim]

x <- as.matrix(seq(0,1,length = n))
```

```
y.DLS <- as.matrix(e.DLS[,1])
plot(x,y.DLS, xlab = " ", ylab = " ")

fit.DLS <-  locpoly(x,y.DLS,bandwidth = dpill(x,y),degree = 1)
plot(fit.DLS, type = "l", xlab = " ", ylab = " ")


y.DDLASSO <- as.matrix(e.DDLASSO[,1])
plot(x,y.DDLASSO, xlab = " ", ylab = " ")

fit.DDLASSO <-  locpoly(x,y.DDLASSO,bandwidth = dpill(x,y),degree = 1)
plot(fit.DDLASSO, type = "l", xlab = " ", ylab = " ")
```

جدول ۱.۴

```
library(lasso2)
data("Prostate")
summary(Prostate)
```

جدول‌های ۲.۴ و ۳.۴

```
library(lasso2)
library(DLASSO)
library(glmnet)
library(msgps)
library(ncvreg)

data("Prostate")
summary(Prostate)

attach(Prostate)
prostate = as.data.frame(Prostate)
prostate[,1:8] = scale(prostate[,1:8])
```

```
prostate[,9] = scale(prostate[,9],scale = FALSE)

x = as.matrix(prostate[,-9])
y = as.matrix(prostate[,9])

dLasso = dlasso(x,y, adp = FALSE, s = 0.001)
AIC.dlasso = dLasso[which(dLasso[,3] == min(dLasso[,3])),3]
BIC.dlasso = dLasso[which(dLasso[,5] == min(dLasso[,5])),5]
AIC.dlasso
BIC.dlasso
DLASSO = coef(dLasso)[,2]
DLASSO

fit.lasso <- glmnet(x, y, family="gaussian", alpha=1)
fit.lasso.cv <- cv.glmnet(x, y, type.measure="mse", alpha=1,
family="gaussian")
lasso = coef(fit.lasso,s = fit.lasso.cv$lambda.min)
lasso



tLL <- fit.lasso$nulldev - deviance(fit.lasso)
k <- fit.lasso$df
n <- fit.lasso$nobs
AICc <- -tLL+2*k+2*k*(k+1)/(n-k-1)
min(AICc)

BIC<-log(n)*k - tLL
min(BIC)

y = as.vector(y)
enet=msgps(x,y,penalty="enet",alpha=0.001)
summary(enet)
```

```
enet=coef(enet)[-1,4]
enet

scad = cv.ncvreg(x,y,penalty = "SCAD")
SCAD<-ncvreg(x,y,penalty = "SCAD",lambda = scad$lambda.min)$beta[-1]
summary(scad)
SCAD
```

جدول‌های ۴.۴ و ۵.۴

```
library(lasso2)
library(DLASSO)
library(glmnet)
library(msgps)
library(ncvreg)

data("Prostate")
summary(Prostate)

attach(Prostate)
prostate = as.data.frame(Prostate)
prostate[,1:8] = scale(prostate[,1:8])
prostate[,9] = scale(prostate[,9],scale = FALSE)

x = as.matrix(prostate[,-9])
y = as.matrix(prostate[,9])
dLasso = dlasso(x,y, adp = FALSE, s = 1)
AIC.dlasso = dLasso[which(dLasso[,3] == min(dLasso[,3])),3]
BIC.dlasso = dLasso[which(dLasso[,5] == min(dLasso[,5])),5]
AIC.dlasso
BIC.dlasso
DLASSO = coef(dLasso)[,2]
DLASSO
```

```
ridge=cv.glmnet(x,y,alpha = 0)
ridge = coef(ridge,s = ridge$lambda.min)[-1]

tLL <- ridge$nulldev - deviance(ridge)
k <- fit.lasso$df
n <- fit.lasso$nobs
AICc <- -tLL+2*k+2*k*(k+1)/(n-k-1)
min(AICc)

BIC<-log(n)*k - tLL
min(BIC)

ridge
```

جدول‌های ۴.۴ و ۵.۴

```
library(lasso2)
library(DLASSO)
library(glmnet)
library(msgps)
library(ncvreg)

data("Prostate")
summary(Prostate)

attach(Prostate)
prostate = as.data.frame(Prostate)
prostate[,1:8] = scale(prostate[,1:8])
prostate[,9] = scale(prostate[,9],scale = FALSE)

x = as.matrix(prostate[,-9])
y = as.matrix(prostate[,9])
```

```
dLasso = dlasso(x,y, adp = FALSE, s = 100)
AIC.dlasso = dLasso[which(dLasso[,3] == min(dLasso[,3])),3]
BIC.dlasso = dLasso[which(dLasso[,5] == min(dLasso[,5])),5]
AIC.dlasso
BIC.dlasso
DLASSO = coef(dLasso)[,2]
DLASSO


ols = lm(y ~ x -1)
ols
AIC(ols)
BIC(ols)
```

جدول ۸.۴

```
setwd("E:/R")
library(haven)
library(latex2exp)
library(DLASSO)
library(msgps)

devtools::install_github('stefano-meschiari/latex2exp')
mroz <- read_dta("mroz.dta")
attach(mroz)
data = as.data.frame(subset(mroz, inlf == 1))
data$wc <-  ifelse(data$educ > 12,1,0)
data$hc <-  ifelse(data$huseduc > 12,1,0)
attach(data)
data.new <- subset(data, select= c(hours,age,nwifeinc,kidslt6,kidsge6,
exper,unem,mtr,wc,hc))
```

```
n <- nrow(data.new)
p <- ncol(data.new)-2


d = c(0.9302,-0.1965,-0.1728,-0.1506,-0.1299,-0.1107,-0.0930,-0.0768)

m = length(d) -1
I.nm <- diag(1,n-m,n-m)
D = matrix(0,nr = (n-m), nc = n)
for(i in 1:(n-m)){
D[i,i:(i+m)] <- d
}
y = data.new$hours
X = as.matrix(data.new[,-c(1,3)])
y.D = D %*% y
X.D = D %*% X
colnames(X.D) <- c("age","k5","k618","exper","unem","mtr","wc","hc")
colnames(y.D) <- "hours"


B.D = solve(t(X.D)%*%M %*% X.D)%*%t(X.D)%*%M %*%y.D

B.LSO = msgps(X.D,y.D,alpha = 0)
B.LSO=coef(B.LSO)[-1,4]

dLasso = dlasso(X.D,y.D, adp = FALSE, s = 0.001)
B.DLSO = coef(dLasso)[,2]


d = c(0.9200,-0.2238,-0.1925,-0.1635,-0.1369,-0.1926,-0.0906)
# d = c(0.9302,-0.1965,0.1728,-0.1506,-0.1299,-0.1107,-0.0930,-0.0768)
m = length(d) -1
```

```
alpha=0.05
N.cv<-500
k<-10 #k-fold cross validation
PE.D.mat=PE.LSO.mat=PE.DLSO.mat=matrix(0,nc=k,nr=N.cv)
for(j in 1:N.cv){
data.new$id<-rep(1:k,length=n)
data.new$id<-sample(data.new$id)
list <- 1:k
for(i.cv in 1:k){
train<-subset(data.new, id %in% list[-i.cv])
test<-subset(data.new, id %in% list[i.cv])
X.train = as.matrix(train[,-c(1,3,11)])
y.train<-as.matrix(train[,1])
n.train <- nrow(X.train)
I.train.nm <- diag(1,n.train-m,n.train-m)
D.train = matrix(0,nr = (n.train-m), nc = n.train)
for(i in 1:(n.train-m)){
D.train[i,i:(i+m)] <- d
}
y.train.D = D.train %*% y.train
X.train.D = D.train %*% X.train
colnames(X.train.D) <- c("age","k5","k618","exper","unem","mtr","wc","hc")
colnames(y.train.D) <- "hours"

X.test<-as.matrix(test[,-c(1,3,11)])
y.test<-as.matrix(test[,1])
n.test <- nrow(X.test)
D.test = matrix(0,nr = (n.test-m), nc = n.test)
for(i in 1:(n.test-m)){
D.test[i,i:(i+m)] <- d
}
y.test.D = D.test %*% y.test
```

```
X.test.D = D.test %*% X.test
colnames(X.test.D) <- c("age","k5","k618","exper","unem","mtr","wc","hc")
colnames(y.test.D) <- "hours"


p = ncol(X.train.D)
M.train =  I.train.nm
- X.train.D %*% solve(t(X.train.D)%*%X.train.D)%*%t(X.train.D)



B.D = solve(t(X.train.D)%*%M.train %*% X.train.D)%*%t(X.train.D)%*%M.train
 %*%y.train.D

PE.D.mat[j,i.cv]<-drop(t(X.test.D%*%B.D -y.test.D)%*%(X.test.D%*%B.D -y.test.D))

y.train.D <- as.vector(y.train.D)
B.LSO = msgps(X.train.D,y.train.D,alpha = 0)
B.LSO=coef(B.LSO)[-1,4]

PE.LSO.mat[j,i.cv]<-t(X.test.D%*%B.LSO -y.test.D)
%*%(X.test.D%*%B.LSO -y.test.D)


dLasso = dlasso(X.train.D,y.train.D, adp = FALSE, s = 0.001)
B.DLSO = coef(dLasso)[,2]
PE.DLSO.mat[j,i.cv]<-t(X.test.D%*%B.DLSO -y.test.D)%*%(X.test.D%*%B.DLSO - y.test.
}
}
PE.D=apply(PE.D.mat,1,mean)
PE.LSO.mat=apply(PE.LSO.mat,1,mean)
PE.DLSO=apply(PE.DLSO.mat,1,mean)
```

```
result1=cbind(mean(PE.D),mean(PE.LSO),mean(PE.DLSO))
result2=cbind(sd(PE.D),sd(PE.LSO),sd(PE.DLSO))/sqrt(N.cv)
result=rbind(result1,result2)
colnames(result)<-c("FM","SM","PT","JS","PR")
rownames(result)<-c("mean","sd")
result[,1]/result
```

مراجع

[۱] الوین، ا.(۱۹۳۴) (ترجمه حسنعلی آذرنوش و ابوالقاسم بزرگ نیا)، *مدل‌های خطی برای آمار* ، د انشگاه فردوسی مشهد (۱۳۸۲) شماره ۳۶۴.

[۲] آرست، م.(۱۳۹۵). *مقایسه رفتار برخی برآوردگرهای انقباضی بریج در مدل رگرسیون چندگانه*، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود.

[۳] برزویی بیدگلی، م.(۱۳۹۳). *بررسی رفتار برآوردگر جک‌نایف ریج*، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود.

[۴] بغلانی، م.(۱۳۹۴). برآوردگرهای تفاضلی استوار در مدل‌های رگرسیون نیمه پارامتریک، پایان‌نامه کارشناسی ارشد، دانشگاه سمنان.

[۵] توماس، ج.، فینی، ر.(۱۹۸۸) (ترجمه مهدی بهزاد، سیامک کاظمی، علی کافی)، *حساب دیفرانسیل و انتگرال و هندسه تحلیلی* تهران، مرکز نشر دانشگاهی (۱۳۷۳-۱۳۷۴).

[۶] روزبه، م.(۱۳۹۰) برآورد در مدل‌های خطی جزئی، رساله دکتری، دانشگاه فردوسی مشهد.

[۷] گلپایگانی، م.(۱۳۹۵) مدل رگرسیون نرخ خطر کاکس جریمه‌شده، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود.

[۸] میرزائی، م.(۱۳۹۴)، مطالعه روشهای تعیین پهنای باند در برآورد کرنل، پایان نامه کارشناسی ارشد دانشگاه صنعتی شاهرود.

[۹] نجاریان، س. (۱۳۹۰). *بررسی رفتار برآوردگرهای رگرسیونی ریج در مدل‌های خطی منفرد*، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود.

[۱۰] نوروزی‌راد، م.(۱۳۹۶) *برآوردگرهای بهبودیافته در برخی مدل‌های رگرسیونی خطی جریمه‌شده*، رساله دکتری، دانشگاه صنعتی شاهرود.

[۱۱] نیرومند، ح. ع.(۱۳۸۷) *تحلیل رگرسیون خطی: ابزاری برای تحقیق*، نشر ارسلان، مشهد.

[12] Abramowitz, M. and Stegun, I. A. (1965), **Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical ables**, Dover Publications. New York.

[13] Ahn, H. and Powell, J.(1993), **Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism**, Journal of Econometrics. **58**, 3-29.

[14] Akaike, H.(1973), **Information Theory and an Extension of the Maximum Likelihood Principle, Second International Symposium on Information Theory**, 267-281.

[15] AlNasser H.,(2014),**On Ridge Regression and Least Absolute Shrinkage and Selection Operator**. University of Victoria.

[16] Arashi, M., Roozbeh, M. and Niroumand, H. (2012), **A Note on Stein-type Shrinkage Estimator in Partial Linear Models**, **Statistics**. **46**, 673-685.

[17] Bai J, and Ng S. (2000). Large dimensional factor analysis. **Foundations and Trends in Econometrics**, **3**(2), 89–163.

[18] Borjesson, P. and Sundberg, C. E. (1979). Simple approximations of the error function $Q(x)$ for communications applications, **IEEE Transactions on Communications**, **27**(3), 639-643.

[19] Buhlmann, P. and van de Geer, S. (2011). **Statistics for High-Dimensional Data: Methods, Theory and Applications**, Springer Series in Statistics. Berlin.

[20] Chevillard, S. and Revol, N.(2008), Computation of the error function erf in arbitrary precision with correct rounding. **JD Bruguera et M. Daumas (editeurs): RNC, 8**, 27-36.

[21] Cody, W. J. (1990). Performance evaluation of programs for the error and complementary error functions, **ACM Transactions on Mathematical Software**, **16**(1), 29-37.

[22] Efron, B and Hastie, T and Johnstone, I and Tibshirani, R and others. (2004), Least angle regression, **The Annals of statistics**, **32**(2), 407-499.

[23] Engle, R.F., Granger, C.W.J., Rice, J. and Weiss, A. (1986), **Semi parametric estimates of the relation between ewather and electricity sales**, Journal of American statistical Association. **81** , 310-320.

[24] Fan, J. and Li, R. (2001). **Variable Selection Via Non-Concave Penalized Likelihood and its Oracle Properties**, Journal of the American Statistical Association, **96**(456), 1348-1360.

[25] Fan, J., Lv, J. and Qi, L., (2011). Sparse high-dimensional models in economics, **Annual Review of Economics**, **3**(1), 291–317.

[26] Fan, J. and Peng, H. (2004), Nonconcave penalized likelihood with a diverging number of parameters, **The Annals of Statistics**, **32**(3), 928-961.

[27] Hall, P., Kay, J. W. and Titterington, D. M., (1990). On estimation of noise variance in two-dimensional signal processing. **Advances in Applied Probability**, **23**, 476-495.

[28] Haselimashhadi, H and Vinciotti, V. (2016), A Differentiable Alternative to the Lasso Penalty, **arXiv preprint arXiv:1609.04985**.

[29] Hoerl, A. E. and Kennard, R. W., (1970). Ridge regression: Biased estimation for nonorthogonal problems, **Technometrics**, **12**(1), 55-67.

[30] Kariya, T. and H. Kurata (2004). **Generalized Least Squares**. Wiley series in Probability and Statistics. New York.

[31] Knight, K. and Fu, W. (2000) Asymptotics for Lasso-type estimators. **The Annals of Statistics , 28**, 1356-1378.

[32] Leal, H. V., Sheissa, R. C., Nino, U. F., Reyes, A. S., and Orea, J. S. (2012) High accurate simple approximation of normal distribution integral. **Mathematical Problems in Engineering**, 1-23.

[33] Lee, C. C. (1992). On Laplace continued fraction for the normal integral, **Annals of the Institute of Statistical Mathematics**, **44**(1), 107-120.

[34] Mac Nally, R.(2000), Regression and model-building in conservation biology, biogeography and ecology: the distinction between–and reconciliation of–predictive and explanatory models, **Biodiversity and Conservation**, **9**(5),655-671.

[35] Mallows, C. L.(1973), Some comments on C_p, **Technometrics**, **15**(4), 661-675.

[36] Montgomery, D. C.(2017), **Design and Analysis of Experiments**, John wiley, New Jersey.

[37] Mroz, T. A., (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. **Econometrica**, **55**(4), 765-799.

[38] Nadaraya, E.A.(1964), On estimating regression, **Theory of Probability and its Applications**. **9** , 141-142.

[39] Nardi, Y. and Rinaldo, A. (2011). **Autoregressive process modeling via the lasso procedure**, Journal of Multivariate Analysis, **102**(3), 528-549.

[40] Press, W.H. (1992). **Numerical Recipes in C: The Art of Scientific Computing**. Number v. 4. Cambridge University Press.

[41] Olver, F.W.J. (2010). National Institute of Standards, and Technology (U.S.). **NIST Handbook of Mathematical Functions**. Cambridge University Press.

[42] Raheem, S. M. E., Ahmed, S. E. and Doksum, K. A. (2012). Absolute penalty and shrinkage estimation in partially linear models, **Computational Statistics and Data Analysis**, **56**, 874-891.

[43] Rice, J. (1984), Bandwidth Choice for Nonparametric Regression, **Annals of Mathematical Statistics**, **12** , 1215-1230.

[44] Roozbe, M., Arashi, M. and Niroumand, H. (2010), **Ridge Regression Methodology in Partial Linear Models with Correlated Errors**, Journal of Statistical Computation and Simulation. **81** , 517-528.

[45] Schmidt, M., Fung, G., and Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. **In Machine Learning: ECML 2007**, pages 286-297.

[46] Schwarz, G. (1978), Estimating the dimension of a model, **The Annals of Statistics**, **6**(2), 461-464.

[47] Stackelberg, H. v. and Heinrich. (1952), **Theory of the Market Economy** Oxford University Press.

[48] Stamey, T. A., Kabalin, J. N., McNeal, J. E. and Johnstone, I. M. and Freiha, F. and Redwine, E. A. and Yang, N.(1989). **Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate. II. Radical Prostatectomy Treated Patients**, The Journal of Urology, **141**(5), 1076–1083.

[49] Tabakan, G. and Akdeniz, F.(2010), **Difference-based Ridge Estimator of Parameters in Partial Linear Model**, Statical Papers. **51**, 357-368.

[50] Tibshirani R.(1996). **Regression Shrinkage and Selection Via the Lasso**, Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[51] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). **Sparsity and Smoothness Via the Fused Lasso**, Journal of the Royal Statistical Society: Series B (Statistical Methodology),**67**(1), 91-108.

[52] Wand, M.P. and Jones, M.C.(1995), **Kernel Smoothing**, Chapman and Hill. London.

[53] Wang, H., Li, G., and Jiang, G. (2007). **Robust Regression Shrinkage and consistent variable Selection Through the LAD-LASSO**, Journal of Business and Economic Statistics, **25**(3), 347-355.

[54] Watson, G.S. (1964), **Smooth regression analysis**, Journal of Sankhya. **26**, 35-72.

[55] van Wieringen, Wessel N. (2015), Lecture notes on ridge regression, **arXiv preprint arXiv:1509.09169**.

[56] Variyan, H. R. (2014). **Big Data: New tricks for Econometrics**, Journal of Economic Perspectives, **28**(2), 3-28.

[57] Yatchew, A. (1997), **An Elemantary Estimator of the Partial Linear Model**, Journal of Economics Letters. **57**, 135-143.

[58] Yatchew, A. (1998), **Nonparametric Regression Techniques in Economics**, Journal of Economic Literature. **36** , 669-721.

[59] Yatchew, A. (1999), **An Elementary Nonparametric Differencing Test of Equality of Regression Functions**, Journal of Economics Letters. **62**, 271-278.

[60] Yatchew, A. (2000), **Scale Economies in Electricity Distribution**, Journal of Applied Economics. **15**, 187-210.

[61] Yatchew, A. (2003), **Semiparametric Regression for the Applied Econometricians**, Cambridge University Press, Cambridge.

[62] 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**(1), 49-67.

[63] Zou, H. (2006), **The adaptive lasso and its oracle properties**, Journal of the American Statistical Association, **101**(476),1418-1429.

[64] 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.

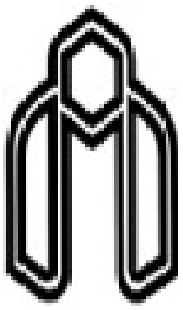
CCC

Abstract

In regression analysis if the number of explanalory variables is lange. estimation of parameters with least squares method is not relevant and one must use penalized estimators.

On of the most efficient penalized estimators is the LASSO. This estimator is not differentiable at zero and hence not normally distributed. In this dissertation we study the performance of a differentiable LASSO which has asymptotic normal distribution. whit some simutation studies and real data analyses, we evaluate the performanc of this estimator along with a difference-based differentiable LASSO.

Keywords:Differentiable LASSO Estimator, Difference Estimator,Penalty Function, High Dimensional Penalized Regression



Shahrood University of Technology

Faculty of Mathematical Sciences

M.Sc. Thesis in Mathematical Statistics

High Dimensional Penalized Regression With Differentiable LASSO

By: Samane Rajabloo

Supervisor

Mohammad Arashi

September 2018