

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

حمایت از حقوق پدیدآورندگان

پایان نامه حاضر، حاصل پژوهشهای نگارنده در دوره کارشناسی ارشد رشته آمار گرایش آمار محض است که در ۱۳۹۳ در دانشکده علوم دانشگاه یاسوج به راهنمایی دکتر آرش اردلان و مشاوره دکتر حیدرعلی مردانی فرد از آن دفاع شده است و کلیه حقوق مادی و معنوی آن متعلق به دانشگاه یاسوج است.



دانشکده علوم
گروه ریاضی

پایان نامه کارشناسی ارشد رشته آمار گرایش آمار محض

رگرسیون لاسو بیزی

استاد راهنما

دکتر آرش اردلان

پژوهشگر

زهرا خادم بشیری

۱۳۹۳



رگرسیون لاسو بیزی

به وسیله

زهرا خادم بشیری

پایان نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی از فعالیت‌های تحصیلی لازم برای اخذ

درجه کارشناسی ارشد

در رشته:

آمار

در تاریخ توسط هیأت داوران زیر بررسی و با درجه به تصویب نهایی رسید.

- | | | | |
|------------------------------------|----------------------------|------------------------|-------|
| ۱- استاد راهنما: | دکتر آرش اردلان | با مرتبه علمی استادیار | امضاء |
| ۲- استاد مشاور: | دکتر حیدرعلی مردانی فرد | با مرتبه علمی استادیار | امضاء |
| ۳- استاد داور داخل گروه: | دکتر کاووس خورشیدیان | با مرتبه علمی دانشیار | امضاء |
| ۴- استاد داور خارج گروه: | دکتر علی‌رضا نعمت‌الهی | با مرتبه علمی استاد | امضاء |
| ۵- نماینده تحصیلات تکمیلی دانشگاه: | دکتر حبیب‌الله خواجه شریفی | با مرتبه علمی استادیار | امضاء |

تقدیم به:

پدر و مادرم

یاوران همیشگی ام

قدردانی

سپاس خداوندگار حکیم را که با لطف بی‌کران خود، آدمی را زیور عقل آراست. در آغاز وظیفه خود می‌دانم از زحمات بی‌دریغ استاد راهنمای خود، جناب آقای دکتر آرش اردلان، صمیمانه تشکر و قدردانی کنم که قطعاً بدون راهنمایی‌های ارزنده ایشان، این مجموعه به انجام نمی‌رسید.

از جناب آقای دکتر حیدرعلی مردانی‌فرد که زحمت مشاوره این رساله را تقبل فرمودند و در آماده‌سازی این رساله، اینجانب را مورد راهنمایی قرار دادند، کمال امتنان را دارم.

در پایان، بوسه می‌زنم بر دستان خداوندگاران مهر و مهربانی، پدر و مادر عزیزم و بعد از خدا ستایش می‌کنم وجود مقدس‌شان را که در این سردترین روزگاران، بهترین پشتیبان من بودند.

زهره خادم بشیری

۱۳۹۳

چکیده

انتخاب بهترین زیرمدل یکی از بحث‌های مهم در مدل‌های رگرسیونی می‌باشد. هدف این روش‌ها این است که پیش‌بین‌های مهم و پیش‌بین‌های قابل اغماض تعیین شده و رابطه‌ی بین متغیر پاسخ و متغیرهای پیش‌بین ساده‌تر بیان شود. علاوه بر این دقت برآوردها و در نتیجه پیش‌بینی مشاهدات آینده نیز افزایش یابد. فرآیندهای انتخاب متغیر کلاسیک از قبیل انتخاب بهترین زیرمجموعه و انتخاب گام به گام، اغلب از لحاظ محاسباتی پرهزینه هستند و گاهی نتایج ناپایداری نیز دارند. بنابراین با توجه به محدودیت‌هایی که روش‌های گام به گام در این زمینه دارند، می‌توان از روش‌های تنظیم بر مبنای رگرسیون جریمه‌دار استفاده کرد. دو روش از روش‌های تنظیم، رگرسیون ستیغی و رگرسیون لاسو است که روش لاسو دارای ویژگی انتخاب متغیر می‌باشد. از آنجا که در بیشتر موارد می‌توان لگاریتم تابع درست‌نمایی را به صورت تابع زیان و تابع چگالی توزیع پیشین را نیز به عنوان تابع جریمه تفسیر کرد، اغلب برای روش‌های تنظیم می‌توان تفسیر بیزی ارائه داد. بنابراین در این رساله لاسو را از دیدگاه بیزی مورد مطالعه قرار داده و اهمیت روش لاسو بیزی را نسبت به روش لاسو معمولی در ایجاد برآورد خطاهای استاندارد و فواصل اعتبار مناسب برای ضرایب رگرسیونی، بیان می‌کنیم.

فهرست مطالب

v	فهرست جداول
vii	فهرست تصاویر
۱	فصل ۱: مقدمه و تاریخچه
۱	فصل ۲: مدل‌های رگرسیونی
۲	۱-۲ مدل رگرسیون خطی ساده
۵	۲-۲ مدل رگرسیون خطی چندگانه
۹	۳-۲ مدل‌های رگرسیونی غیرخطی
۹	۱-۳-۲ رگرسیون چندجمله‌ای
۱۱	۲-۳-۲ نوارهای باریک رگرسیونی
۱۶	۳-۳-۲ مقایسه‌ی رگرسیون چندجمله‌ای و نوارهای باریک
۱۶	۴-۳-۲ انتخاب گره و مکان آن
۱۸	۵-۳-۲ نوارهای باریک هموار
۲۰	۶-۳-۲ مدل‌های جمعی تعمیم یافته
۲۹	۴-۲ رگرسیون جریمه‌دار
۳۰	۱-۴-۲ رگرسیون ستیغی
۳۳	۲-۴-۲ رگرسیون لاسو
۳۵	۳-۴-۲ تفسیر هندسی رگرسیون ستیغی و لاسو
۳۶	۴-۴-۲ حالتی خاص از رگرسیون ستیغی و لاسو
۴۰	فصل ۳: مدل‌های بیزی

۴۱	۱-۳ آمار بیز
۴۴	۱-۱-۳ چگالی پیشین و چگالی پسین
۴۵	۲-۱-۳ تعریف خانواده چگالی مزدوج
۴۶	۳-۱-۳ چگالی پیشین ناآگاهی بخش
۴۷	۴-۱-۳ پیشین جفریز
۴۸	۲-۳ مقدمه‌ای از استنباط آماری به روش بیزی
۵۱	۳-۳ مدل رگرسیون خطی چندگانه بیزی
۵۲	۱-۳-۳ مدل رگرسیون چندگانه بیزی با پیشین توام
۵۵	۲-۳-۳ چگالی پسین حاشیه‌ای از β
۵۷	۳-۳-۳ چگالی‌های پسین حاشیه‌ای برای τ و σ^2
۶۰	فصل ۴: روش های مونت کارلو در مدل‌های بیزی
۶۲	۱-۴ محاسبه ی انتگرال‌ها به روش مونت کارلو
۶۶	۱-۱-۴ تقریب پارامترها براساس نماهای پسین
۶۸	۲-۱-۴ روش برآورد مونت کارلو
۷۱	۳-۱-۴ استخراج نمونه از چگالی f
۸۱	۲-۴ زنجیره های مارکف
۸۳	۱-۲-۴ ساختار احتمالی زنجیره
۸۴	۲-۲-۴ توزیع حدی زنجیره
۸۵	۳-۲-۴ توزیع ایستای زنجیره
۹۲	۴-۲-۴ کاربرد زنجیره های مارکف و ارگودیکی در مسائل بیزی
۹۳	۳-۴ روش های مونت کارلو بر اساس زنجیره های مارکف
۹۴	۱-۳-۴ الگوریتم متروپلیس - هستینگ
۹۶	۲-۳-۴ نمونه‌گیری گیبز
۱۰۱	۴-۴ تحلیل خروجی‌ها در MCMC
۱۱۰	۵-۴ بهینه‌سازی مونت کارلو
۱۱۱	۱-۵-۴ روش‌های بهینه‌سازی عددی
۱۱۲	۲-۵-۴ تقریب تصادفی

۱۱۸	۶-۴ تقریب لاپلاس آشیانه‌ای تلفیقی
۱۲۱	فصل ۵: رگرسیون لاسو بیزی
۱۲۱	۱-۵ توضیح ساده‌ای از لاسو و رگرسیون کمترین زاویه
۱۲۲	۲-۵ محاسبه‌ی برآوردها در رگرسیون لاسو
۱۲۲	۲-۵-۱ الگوریتم رگرسیون حداقل زاویه (LAR)
۱۲۴	۳-۵ تفسیر بیزی لاسو
۱۳۰	۳-۵-۱ فرمول‌بندی مدل سلسله مراتبی
۱۳۳	۳-۵-۲ اجرای نمونه‌گیر گیز
۱۳۶	۳-۵-۳ انتخاب پارامتر در رگرسیون لاسو بیزی
۱۳۹	۴-۵ رگرسیون لاسو بیزی از دیدگاهی متفاوت
۱۴۲	۴-۵-۱ توزیع نرمال متعامد کنج
۱۴۴	۴-۵-۲ پیش‌بینی براساس پسین
۱۵۰	۴-۵-۳ نمونه‌گیر گیز استاندارد
۱۵۱	۴-۵-۴ نمونه‌گیر گیز متعامدشده
۱۵۴	فصل ۶: بحث و نتیجه‌گیری
۱۵۴	۶-۱ برآورد پارامترها به روش رگرسیون گام به گام پیشرو
۱۵۷	۶-۲ برآورد پارامتر به روش رگرسیون ستیغی و لاسو
۱۶۱	۶-۳ برآورد پارامترها در رگرسیون لاسو بیزی با استفاده از الگوریتم <i>lars</i>
۱۶۶	۶-۴ برآورد پارامتر به روش رگرسیون لاسو بیزی
۱۶۶	۶-۴-۱ برآورد پارامترها به روش متروپلیس - هستینگ
۱۷۴	۶-۴-۲ برآورد پارامترها به روش گیز
۱۷۷	۶-۴-۳ برآورد پارامتر لاسو بیزی
۱۷۸	۶-۴-۴ بررسی آمیختگی زنجیره مارکف تولید شده از روش گیز
۱۸۲	۶-۵ پیشنهادات
۱۹۴	پیوست آ: تعاریف و مفاهیم آماری
۱۹۴	آ-۱ پایه

۱۹۶	۲-آ روش‌های بازنمونه‌گیری
۱۹۷	۱-۲-آ روش مجموعه اعتبار
۱۹۸	۲-۲-آ روش اعتبارسنجی متقابل Leave-One-Out
۱۹۹	۳-۲-آ روش اعتبارسنجی متقابل k تایی
۲۰۰	۳-آ متمرکز کردن متغیر کمکی
۲۰۲	۴-آ انتخاب مدل خطی
۲۰۳	۱-۴-آ بهترین انتخاب زیرمجموعه
۲۰۴	۲-۴-آ انتخاب گام به گام پیشرو
۲۰۵	۳-۴-آ انتخاب گام به گام پسرو
۲۰۸	۵-آ توابع زیان
۲۰۹	۶-آ تقریب لاپلاس
۲۱۱	۷-آ ماکزیمم درست‌نمایی حاشیه‌ای
۲۱۲	۸-آ برآورد بیز تجربی
۲۱۵	۹-آ بیز سلسله مراتبی
۲۱۶	۱۰-آ فاصله اطمینان بزرگترین چگالی پسین
۲۱۷	۱۱-آ خطای استاندارد مونت کارلو
۲۱۸	۱۲-آ تقلیل کردن زنجیره
۲۱۹	۱۳-آ تعیین همگرایی زنجیره مارکف

پیوست ب: برنامه‌های نوشته شده برای رسم نمودارها، برازش مدل و برآورد پارامترها

۲۲۰	با استفاده از R
۲۲۰	ب-۱ مدل‌های رگرسیونی
۲۲۰	ب-۱-۱ مدل‌های رگرسیونی غیر خطی
۲۲۱	ب-۱-۲ مدل‌های جمعی تعمیم یافته
۲۲۴	ب-۲ مدل‌های بیزی
۲۲۴	ب-۲-۱ مدل رگرسیون خطی چندگانه بیزی
۲۲۶	ب-۲-۲ روش‌های مونت کارلو در مدل‌های بیزی
۲۲۸	ب-۲-۳ روش برآورد مونت کارلو

- ب-۲-۴ زنجیره‌های مارکف ۲۳۲
- ب-۲-۵ روش‌های مونت کارلو بر اساس زنجیره‌های مارکف ۲۳۳
- ب-۳ رگرسیون لاسو بیزی ۲۳۹

۲۴۱

مراجع

فهرست جداول

- ۴-۱ کد مربوط به الگوریتم رد ۷۶
- ۶-۱ کد مربوط به روش گام به گام پیشرو ۱۵۶
- ۶-۲ کد مربوط به روش گام به گام پیشرو ۱۵۷
- ۶-۳ کد مربوط به برآورد در روش رگرسیون ستیغی ۱۵۸
- ۶-۴ کد مربوط به برآورد خطای آزمون در روش رگرسیون ستیغی ۱۵۹
- ۶-۵ کد مربوط به برآورد ضرایب در روش رگرسیون ستیغی ۱۶۰
- ۶-۶ کد مربوط به برآورد در روش رگرسیون لاسو ۱۶۰
- ۶-۷ کد مربوط به برآورد خطای آزمون و انتخاب متغیر در روش رگرسیون لاسو ۱۶۱
- ۶-۸ کد مربوط به الگوریتم LAR ۱۶۳
- ۶-۹ کد مربوط به الگوریتم LAR ۱۶۳
- ۶-۱۰ کد مربوط به خروجی الگوریتم LAR ۱۶۵
- ۶-۱۱ کد SAS مربوط به فراخوانی داده‌های دیابت ۱۶۷
- ۶-۱۲ کد مربوط به استاندارد کردن متغیرهای مدل با استفاده از SAS ۱۶۷
- ۶-۱۳ کد SAS مربوط به برازش مدل لاسو بیزی ۱۶۸
- ۶-۱۴ ادامه‌ی کد SAS مربوط به برازش مدل لاسو بیزی ۱۸۵
- ۶-۱۵ ادامه‌ی کد SAS مربوط به فواصل اطمینان در برازش مدل لاسو بیزی ۱۸۶
- ۶-۱۶ کد SAS مربوط به خروجی فواصل اعتبار در برازش مدل لاسو بیزی ۱۸۷

- ۱۷-۶ کد مربوط به مدل اول در تابع blasso ۱۸۷
- ۱۸-۶ کد مربوط به مدل دوم در تابع blasso ۱۸۷
- ۱۹-۶ کد مربوط به مدل سوم در تابع blasso ۱۸۸
- ۲۰-۶ کد مربوط به گیز استاندارد و متعامد شده ۱۸۸
- ۲۱-۶ کد مربوط به گیز متعامد شده برای داده‌های دیابت ۱۸۹
- ۲۲-۶ کد مربوط به استنباطها بر اساس پسین برای مدل لاسو بیزی ۱۹۰
- ۲۳-۶ کد مربوط به برآورد پارامتر لاسو بیزی برای داده‌های دیابت ۱۹۱
- ۲۴-۶ کد مربوط به بررسی آمیختگی و همبستگی زنجیره مارکف تولید شده از گیز ۱۹۲
- ۱۲۵-۶ کد ادامه‌ی کد مربوط به بررسی آمیختگی و همبستگی زنجیره مارکف تولید شده از گیز ۱۹۳

فهرست تصاویر

- ۲-۱ فرانسيس گالتون (۱۸۲۲ - ۱۹۱۱) ۲
- ۲-۲ برآزش رگرسيون خطی ساده ۶
- ۲-۳ برآزش رگرسيون خطی ساده ۱۰
- ۲-۴ برآزش رگرسيون خطی چندجمله‌ای ۱۱
- ۲-۵ برآزش رگرسيون چندجمله‌ای تکه‌ای ۱۳
- ۲-۶ مقایسه‌ی برآزش رگرسيون چندجمله‌ای و نوار باریک برای داده‌های fossil ۱۷
- ۲-۷ برآزش نوار باریک درجه ۳ برای داده‌های fossil ۱۸
- ۲-۸ نمایش تابع تک متغیره با استفاده از توابع پایه و نوارهای باریک درجه ۳ ۲۲
- ۲-۹ داده‌های مشاهده شده برای موتور ولوو ۲۴
- ۲-۱۰ برآزش‌های نوار باریک رگرسيوني جریمه‌دار برای داده‌های پوسیدگی ۲۷
- ۲-۱۱ برآزش مدل بهینه با استفاده از GCV ۳۰
- ۲-۱۲ برآورد ضرایب رگرسيون ستیغی برای مجموعه داده‌های Hitter ۳۳
- ۲-۱۳ برآورد ضرایب رگرسيون لاسو برای مجموعه داده‌های Hitter ۳۵
- ۲-۱۴ نمایش هندسی رگرسيون ستیغی و لاسو: در این نمودار، $p = ۲$ و با توجه به شکل سمت راست برخورد در $\beta_1 = ۰$ رخ داده است پس مدل حاصل فقط شامل β_2 است. ۳۶
- ۲-۱۵ مقایسه‌ی رگرسيون ستیغی و لاسو با روش حداقل مربعات ۳۸
- ۳-۱ تصویری از کشیش توماس بیز در سمت چپ و سیمون لاپلاس در سمت راست ۴۰
- ۳-۲ ترکیب اطلاعات توسط قضیه‌ی بیز ۴۳

- ۴-۱ نمودار کانتور چگالی پسین پارامترهای (τ, K) ۶۴
- ۴-۲ نمودار کانتور چگالی پسین پارامترهای (τ) ۶۵
- ۴-۳ نمودار کانتور تقریب نرمال برای پارامترهای $(\tau, \log K)$ در مدل بتا -
 دو جمله‌ای ۶۸
- ۴-۴ تقریب انتگرال تابع $h(x) = [\cos(5 \circ x) + \sin(2 \circ x)]^2$ ۷۱
- ۴-۵ نمودار تولید متغیرهای تصادفی نمایی با استفاده از تبدیل معکوس (سمت راست) و با استفاده از دستور rexp در R (سمت چپ) مشاهده می‌شود.
 منحنی چگالی $\exp(1)$ نیز در بالای این نمودار رسم شده است. ۷۳
- ۴-۶ الگوریتم پذیرش/رد: انتخاب θ از توزیع پیشنهادی و انتخاب یک متغیر تصادفی U از توزیع یکنواخت. قبول نمونه‌ی کاندیدی اگر $U \leq \frac{g(\theta|y)}{cp(\theta)}$
 باشد و در غیر اینصورت رد آن. ۷۴
- ۴-۷ نمودار کانتور مقادیر شبیه‌سازی شده در الگوریتم پذیرش/رد ۷۷
- ۴-۸ تولید متغیرهای تصادفی $x \sim \beta e(2/7, 6/3)$ با استفاده از الگوریتم پذیرش/رد ۷۷
- ۴-۹ نمودار چگالی‌های پیشنهادی، پسین و تابع وزن با تقریب نرمال ۸۰
- ۴-۱۰ نمودار چگالی‌های پیشنهادی، پسین و تابع وزن با تقریب توزیع t ۸۰
- ۴-۱۱ نمودار زنجیره مارکف با ماترس انتقال S ۸۹
- ۴-۱۲ نمودار مربوط به زنجیره مارکف با فضای حالت $1, S = 0$ ۹۱
- ۴-۱۳ نمودار مربوط به زنجیره مارکف با فضای حالت $S = \mathcal{Z}$ ۹۲
- ۴-۱۴ نمودار مربوط به اجرای الگوریتم متروپلیس - هستینگ برای شبیه‌سازی نمونه‌هایی از چگالی نمایی دوگانه ۹۷
- ۴-۱۵ کد مربوط به نمودار کانتور چگالی پسین ۱۰۸
- ۴-۱۶ برآمدهای شبیه‌سازی شده‌ی μ و $\log \sigma$ ، حاصل از الگوریتم قدم تصادفی متروپلیس ۱۰۹
- ۴-۱۷ نمودار خودهمبستگی دنباله‌ی حاصل از الگوریتم قدم تصادفی متروپلیس ۱۱۰
- ۵-۱ نمایش الگوریتم LARS در حالت دو متغیره ۱۲۳
- ۵-۲ مقایسه‌ی چگالی نرمال و لاپلاس ۱۲۷

- ۳-۵ نمودار کانتور توزیع پسین تحت پیشین غیر شرطی ۱۲۸
- ۴-۵ نمایش چگالی‌های پسین تک متغیره برای پنج مقدار مختلف τ به طوری
که نمودار ضخیم مربوط به $\tau = ۲$ می‌باشد. ۱۴۸
- ۵-۵ نمایش توزیع پیش‌بین پسین ۱۴۹
- ۱-۶ انتخاب بهترین مدل براساس R^2 تعدیل یافته در روش گام به گام پیشرو . ۱۵۵
- ۲-۶ نمودار ضرایب برازش رگرسیون ستیغی و لاسو ۱۶۲
- ۳-۶ نمودار ضرایب در الگوریتم lars ۱۶۴
- ۴-۶ آمیختگی زنجیره با اجرای گیبز استاندارد ۱۸۰
- ۵-۶ آمیختگی زنجیره با اجرای گیبز متعامد شده ۱۸۱
- ۱-آ نمایش پایه‌ی مرتبط با مدل خطی مستقیم ۱۹۴
- ۲-آ نمایش مدل شکسته شده در یک نقطه ۱۹۵
- ۳-آ نمایش پایه‌ی مرتبط با مدل شکسته شده در یک نقطه ۱۹۵
- ۴-آ نمایش پایه‌ی مرتبط با مدل دارای چندین شکستگی ۱۹۶

فصل ۱

مقدمه و تاریخچه

برای پیش‌بینی یک متغیر وابسته بر اساس k متغیر مستقل اغلب از مدل رگرسیونی

$$y = \sum_{j=1}^k x_j^T \beta + \varepsilon,$$

استفاده می‌شود. در این مدل، y مربوط به متغیر وابسته، x_j متغیر مستقل، β بردار ضرایب رگرسیون و ε خطای تصادفی است. ضرایب رگرسیون، پارامترهای مدل را تشکیل داده که مقادیر آن‌ها در همه‌ی موارد ثابت است، درحالی‌که ε به طور تصادفی از یک مورد به مورد دیگر تغییر می‌کند. ممکن است در طی مراحل استنباط، قصد داشته باشیم برخی پیش‌بین‌ها را که تاثیر زیاد یا ارتباط زیادی با متغیر پاسخ ندارند را از مدل حذف کرده و استنباط بهبود یابد.

جستجو برای بهترین زیرمدل (یا مجموعه‌ای از زیرمدل‌ها) را انتخاب متغیر^۱ یا انتخاب زیرمجموعه^۲ می‌نامند.

در تحلیل رگرسیونی، انتخاب متغیر نقش مهمی در استخراج اطلاعات از مجموعه داده‌های بزرگ با ساختارهای پیچیده دارد. باید توجه داشت که انتخاب تعداد زیاد متغیرهای توضیحی، واریانس مدل ایجادشده را افزایش داده و تعداد کم متغیرها نیز منجر به برآوردهای ناسازگار خواهد شد.

^۱ Variable Selection

^۲ Subset Selection

روش‌های کلاسیک زیادی برای انتخاب متغیر وجود دارد که برخی از آن‌ها مانند رگرسیون گام به گام^۳ براساس دنباله‌ای از آزمون‌های فرض و یا برآوردهایی از نوع میانگین توان دوم خطا یا دیگر معیارهای زیان می‌باشند. این روش‌ها بیشتر بر روی مسئله‌ی انتخاب متغیر متمرکز شده و به برآورد ضرایب نمی‌پردازند، همچنین در این روش‌ها، تغییرات کوچک در داده‌ها موجب می‌شود یک متغیر به جای دیگری انتخاب شده و نتایج انتخاب‌ها به‌طور کامل متفاوت شود. به علاوه این روش‌ها از تکنیک حداقل مربعات برای برازش یک مدل خطی که شامل زیرمجموعه‌ای از پیش‌بین‌هاست، استفاده می‌کنند.

به عنوان یک جایگزین انتخاب مدل، می‌توان از روش‌های تنظیم^۴ استفاده کرد. این روش‌ها در مواردی که تعداد پیش‌بین‌ها زیاد و یا حتی بیشتر از تعداد مشاهدات است و روش‌های تنظیم نشده منجر به بیش‌برازشی می‌شود کاربرد دارند.

اخیراً دو روش رگرسیون ستیغی^۵ و رگرسیون لاسو^۶ به عنوان روش‌های انتخاب متغیر مورد توجه آمادانان قرار گرفته است. این روش‌ها، همچنین به روش‌های انقباض^۷ نیز مشهورند. در این روش‌ها برآوردهای ضرایب محدود شده و به سمت صفر کاهش می‌یابند که این کاهش می‌تواند به‌طور معناداری واریانس ضرایب را کاهش دهد. رگرسیون ستیغی را هورل و کنارد^۸ در سال ۱۹۷۰ [۱] ارائه کردند. در این روش ضرائب به صفر میل کرده و از آنجایی که مقادیر آن دقیقاً صفر نمی‌شوند، از مدل به‌طور کامل حذف نمی‌شوند اما اثر آن‌ها بسیار کم می‌شود. بنابراین این روش را نمی‌توان برای هدف انتخاب متغیر استفاده کرد.

روش لاسو که در سال ۱۹۹۶ توسط تیشیرانی^۹ [۲] ارائه شده، به عنوان روشی برای انتخاب مدل و برآورد پارامتر به‌طور همزمان، مورد توجه قرار گرفته است. مزایای اصلی و مهم لاسو ترکیبی از افزایش دقت پیش‌بینی و بهبود تفسیر مدل‌های ساخته شده می‌باشند.

برآوردهای لاسو کاربردهای ساده‌ای از روش‌های برآورد پارامتر (مانند حداقل مربعات

Stepwise Regression^۳Regularization^۴Ridge Regression^۵Lasso Regression^۶Shrinkage^۷Hoerl and Kennard (1970)^۸Tibshirani(1996)^۹

و روش گشتاورها) را اجرا کرده و ضرایب با پیش‌بین‌های ناچیز را دقیقاً به صفر کاهش می‌دهند. بنابراین، مدل‌های حاصل، روی اثرگذارترین اثرات، متمرکز شده و باعث دقت پیش‌بینی می‌گردد. علاوه بر این، برآوردهای لاسو پایتر از دیگر روش‌های انتخاب متغیر که براساس دیگر معیارها از قبیل AIC و غیره هستند، می‌باشند. مزیت دیگر روش لاسو، از لحاظ محاسباتی است. چون هزینه‌ی محاسبات آن به سختی فراتر از پیچیدگی رگرسیون خطی می‌باشد. همچنین روش محاسباتی آن ساده‌تر از دیگر روش‌های انتخاب متغیر کلاسیک بوده و نیازمند جستجوهای ترکیبی سخت نمی‌باشد.

علاوه بر مزایایی که روش لاسو دارد، این برآوردگر محدودیت‌هایی نیز داشته و در مواردی که پیش‌بین‌ها همخطی بالایی دارند، نتایج ناسازگاری خواهد داشت. برای حل این مشکل در سال ۲۰۰۵، زو و هستی^{۱۰} [۳] روش شبکه ارتجاعی^{۱۱} را ارائه دادند که در آن ترکیبی از جریمه‌ی رگرسیونی ستیغی و لاسو استفاده شده است. در سال ۲۰۰۶ نیز زو^{۱۲} [۴]، روش لاسو سازوار^{۱۳} را معرفی کرده که در این روش مقادیر مختلف انقباض برای هر ضریب رگرسیونی اعمال می‌شود.

در همه‌ی این روش‌ها، برآوردهایی که برای پارامترها به دست می‌آیند، برآوردهایی نقطه‌ای هستند. همان‌طور که می‌دانیم برآوردهای فاصله‌ای نسبت به برآورد نقطه‌ای ارجحیت داشته و با یک میزان اطمینان مشخص می‌شوند، در صورتی که در برآورد نقطه‌ای مقداری را به‌عنوان برآورد انتخاب می‌کنیم که احتمال آن صفر است. در سال ۲۰۰۰، نایت و فو^{۱۴} [۵] تقریب‌هایی را برای برآوردهای لاسوگونه در نظر گرفتند که برای ایجاد مجموعه فواصل اطمینان فراوانی‌گرا مورد استفاده قرار گرفت ولی در آن ویژگی‌های موردنیاز برای موارد با نمونه کوچک، واضح و روشن نبود.

مسئله‌ی دیگری که در لاسو مد نظر است مربوط به برآورد خطاهای استاندارد برآورد پارامترها می‌باشد، زیرا الگوریتم‌های موجود از قبیل الگوریتم LARS فقط برآوردهای نقطه‌ای پارامترها

Hastie(۲۰۰۵) and^{۱۰}Elastic Net^{۱۱}Zou(2006)^{۱۲}Adaptive Lasso^{۱۳}Knight and Fu(2000)^{۱۴}

را فراهم می‌کنند. همچنین استفاده از بوت استرپ برای محاسبه‌ی این برآوردها از لحاظ محاسباتی مشکل است. علاوه بر این‌ها، ساختار سلسله مراتبی مسئله در چهارچوب فراوانی‌گرا قابل حل نبوده و نیازمند روش‌های بیزی است.

روش‌های بیزی برآوردهای فاصله‌ای را برای ما فراهم کرده و در حالت‌هایی که تعداد مشاهدات کم و پارامترها زیاد است، عملکرد خوبی دارند. این روش‌ها در فرآیند انتخاب متغیر نیز بهتر از روش‌های غیربیزی عمل می‌کنند.

در سال ۱۹۸۸ میچل و بوچمپ [۶] ^{۱۵} روش انتخاب متغیر بیزی را برای رگرسیون خطی ارائه دادند. در این روش، یک توزیع احتمال (توزیع پیشین) به ضرایب نسبت داده شده و در آن توزیع هر β_j ای که مرتبط با یک پیشین آسیب‌پذیر است شامل جرم احتمالی گسسته در نقطه‌ی $\beta_j = 0$ می‌باشد.

پیشینی که آن‌ها برای هر پارامتر در نظر گرفتند، پیشینی با چگالی Spike and Slab بود. چگالی پیشین به این صورت است که در آن β_j به‌طور یکنواخت بین دو محدودیت $-f_j$ و f_j توزیع شده و فقط مواقعی که x_j در معرض حذف است، یک بخش از جرم احتمال در صفر متمرکز می‌شود.

سپس، با استفاده از قضیه‌ی بیز احتمالات پسین محاسبه شده و از این احتمالات برای یافتن بهترین زیرمدل یا مجموعه‌ای از زیرمدل‌ها استفاده می‌شود. محدودیت‌های عملی این روش شامل محاسبات پیچیده در حالت زیادی تعداد پیش‌بین‌ها، محروم‌سازی کلاسی از متغیرها با بیش از دو سطح و کنار گذاشتن وابستگی‌های تابعی بین پیش‌بین‌ها می‌باشد.

در سال ۲۰۰۵ یان و لین [۷] ^{۱۶}، پیشینی برای ضرایب رگرسیونی در نظر گرفتند که شامل یک Spike تهایده در صفر و Slab نمایی دوگانه بوده و سپس با استفاده از آن تحلیل بیزی برای تقریب پسین را اجرا کردند. البته تحلیل آن‌ها برآوردهایی مشابه لاسو اصلی و بدون هیچ برآورد فاصله‌ای مرتبطی را نتیجه داد.

مدل رگرسیون لاسو بیزی نیز یکی دیگر از روش‌های انتخاب مدل است که در سال ۱۹۹۶ توسط تیشیرانی ارائه و در سال ۲۰۰۸ توسط پارک و کسلا [۸] ^{۱۷} اجرا شد. اما روشی که

^{۱۵} Mitchell and Beauchamp(1988)

^{۱۶} Yuan and Lin(2005)

^{۱۷} Park and Casella(2008)