حمایت از حقوق پدیدآورندگان

پایان نامهٔ حاضر، حاصل پژوهشهای نگارنده در دورهٔ **کارشناسی ارشد** رشتهٔ **آمار** گرایش **آمار محض** است که در **۱۳۹۳** در **دانشکده علوم** دانشگاه **یاسوج** به راهنمایی **دکتر آرش اردلان** و مشاورهٔ **دکتر حیدرعلی مردانی فرد** از آن دفاع شده است و کلیهٔ 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 وسیلهٔ

زهرا خادم بشیری

پایان نامه

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

درجهٔ کارشناسی ارشد

در رشتهٔ:

آمار

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

۱- استاد راهنما:	دکتر آرش اردلان	با مرتبهٔ علمی استادیار	امضاء
۲- استاد مشاور:	دکتر حیدرعلی مردانی فرد	با مرتبهٔ علمی استادیار	امضاء
۳- استاد داور داخل گروه:	دکتر کاووس خورشیدیان	با مرتبهٔ علمی دانشیار	امضاء
۴- استاد داور خارج گروه:	دکتر علی‌رضا نعمت‌الهی	با مرتبهٔ علمی استاد	امضاء
۵- نمایندهٔ تحصیلات تکمیلی دانشگاه:	دکتر حبیب الله خواجه شریفی	با مرتبهٔ 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.

زهرا 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 گام‌به‌گام 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 ارایه 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

فهرست جداول

فهرست تصاویر

فصل ۱

مقدمه و تاریخچه

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

$$\mathrm{y} = \sum_{j=1}^{k} \mathrm{x}_j^T \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

استفاده می‌شود. در این مدل، y مربوط به متغیر وابسته، x_j متغیر مستقل، $\boldsymbol{\beta}$ 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 بیش برازشی می‌شود کاربرد دارند.

اخیرا دو روش رگرسیون ستیغی [۵] و رگرسیون لاسو [۶] به عنوان روش‌های انتخاب متغیر مورد توجه آماردانان 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 دقیقا صفر نمی‌شوند، از مدل به طور 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 تیبشیرانی [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.

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

---

[۳] Stepwise Regression

[۴] Regularization

[۵] Ridge Regression

[۶] Lasso Regression

[۷] Shrinkage

[۸] Hoerl and Kennard (1970)

[۹] Tibshirani(1996)

و روش گشتاورها) را اجرا کرده و ضرایب با پیش‌بین‌های ناچیز را دقیقا به صفر کاهش می‌دهند. بنابراین، مدل‌های حاصل، روی اثرگذارترین اثرات، متمرکز شده و باعث دقت پیش‌بینی می‌گردد. علاوه بر این، برآوردهای لاسو پایاتر 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 صورتی‌که 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

---

[۱۰] Hastie(۲۰۰۵) and

[۱۱] Elastic Net

[۱۲] Zou(2006)

[۱۳] Adaptive Lasso

[۱۴] Knight and Fu(2000)

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_jای که مرتبط با یک پیشین آسیب‌پذیر است شامل جرم احتمالی گسسته در نقطه‌ی $\beta_j = ۰$ 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 تباهیده 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 تیبشیرانی 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)