دانشکده علوم
گروه ریاضی

پایان نامه کارشناسی ارشد رشته آمار گرایش آمار ریاضی

رگرسیون با انقباض نقاط دورافتاده

استاد راهنما

دکتر حیدرعلی مردانی‌فرد

پژوهشگر

شهره شریفی میناب

بهمن ماه ۱۳۹۳

حمایت از حقوق پدیدآورندگان

دانشکده علوم
گروه ریاضی

پایان نامه کارشناسی ارشد رشته آمار گرایش آمار ریاضی

رگرسیون با انقباض نقاط دورافتاده

استاد راهنما

دکتر حیدرعلی مردانی‌فرد

پژوهشگر

شهره شریفی میناب

بهمن ماه ۱۳۹۳

رگرسیون با انقباض نقاط دورافتاده

به وسیلهٔ

شهره شریفی میناب

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 حیدرعلی مردانی‌فرد	با مرتبهٔ علمی استادیار	امضاء
۲- استاد مشاور:	دکتر آرش اردلان	با مرتبهٔ علمی استادیار	امضاء
۳- استاد داور داخل گروه:	دکتر روح الله روزگار	با مرتبهٔ علمی استادیار	امضاء
۴- استاد داور خارج گروه:	دکتر علیرضا نعمت الهی	با مرتبهٔ علمی استاد	امضاء
۵- نمایندهٔ تحصیلات تکمیلی دانشگاه:	دکتر حمیدرضا رجبی	با مرتبهٔ علمی استادیار	امضاء

تقدیم به:

پیشگاه مقدس مولای برحق عالم حضرت مهدی(عج)

و

خانواده مهربانم

قدردانی

سپاس و ستایش مر خدای را جل جلاله که آثار قدرت او بر چهره روز روشن، تابان است و انوار حکمت او در دل شب تار، درفشان. آفریدگاری که خویشتن را به ما شناساند و درهای علم را برما گشود و عمری و فرصتی عطا فرمود تا بدان بنده ضعیف، خویشتن را در طریق علم و معرفت بیازماید.

با درود فراوان به روح پرفتوح پدرم و سپاس بیکران بر همدلی و همراهی و همگامی مادر دلسوز و مهربانم که سجده‌ی ایثارش گل محبت را در وجودم پروراند و دامان گهربارش لحظه‌های مهربانی را بر من آموخت و شکر خدای را از وجود برادر و خواهرانی عزیز که اگر نبود دستان یاری گرشان این معنا قطعا محقق نمی‌گشت.

در این مجال بر خود واجب و لازم می‌دانم از استاد باکمالات و شایسته، جناب آقای **دکتر حیدرعلی مردانی‌فرد** که در نهایت صبوری و مهربانی، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این پروژه را برعهده گرفتند، همچنین از استاد گرانقدر، جناب آقای **دکتر آرش اردلان** مدیریت محترم کرسی گروه، که زحمت مشاوره این رساله را در حالی متقبل شدند که بدون مساعدت ایشان، این پروژه به نتیجه نمی‌رسید و بعلاوه از اساتید فرزانه و عالم جناب آقایان **دکتر علیرضا نعمت الهی** و **دکتر روح الله روزگار** که زحمت داوری رساله را متقبل شدند کمال تشکر و قدردانی را دارم.

باشد که این خردترین، بخشی از زحمات آنان را سپاس گوید.

شهره شریفی میناب

بهمن ماه ۱۳۹۳

چکیده

در آمار از جمله ابزار مهم برای تحلیل داده‌ها برآورد مناسب یک تابع است که روش‌های مختلفی برای آن در حالت‌های پارامتری و ناپارامتری ارائه شده است. یکی از معروف‌ترین روش‌ها در برآورد توابع پارامتری، روش کمترین توان‌های دوم عادی است که در شرایط مطلوب از مزیت‌های زیادی برخوردار است. با این وجود یک نقطه ضعف بسیار مهم این روش تاثیرپذیری آن از نقاط دورافتاده‌ای است که خواسته یا ناخواسته در مجموعه‌ی مشاهدات حضور پیدا می‌کنند. ایده‌ی استفاده از رگرسیون استوار بر این اساس شکل گرفته است که در آن تاثیر نقاط دورافتاده را کاهش داده و به روش بکار رفته اجازه برآورد دقیق‌تر پارامترها را بدهد.

علاوه بر این گاهی به علت حضور تعداد زیاد متغیرهای پیش‌بین در مدل، تفسیر آن دشوار خواهد بود. در این مواقع محقق سعی می‌کند تعداد متغیرهای پیش‌بین را کاهش داده و زیرمجموعه‌ای از متغیرها در بین تمام پیش‌بین‌ها انتخاب کند. یکی از روش‌های موثر در این زمینه استفاده از رگرسیون تاوانیده است که تاثیر آن بر اندازه پارامترها و میزان تمایل آنها به صفر می‌باشد.

در این راستا با ترکیب هردو روش استوار و تاوانیده قادر به ارائه روش جدیدی با نام "رگرسیون با انقباض نقاط دورافتاده" هستیم که هم نسبت به روش‌های دیگر از استواری بیشتری برخوردار است وهم با تعمیم آن به "رگرسیون تنک با انقباض نقاط دورافتاده" قادر به انتخاب متغیر و همچنین افزایش استواری مدل خواهیم بود که از آسانی بیشتری در محاسبات و کاربرد برخوردار است. پس از معرفی این برآوردگرها به مقایسه آنها با دیگر مدل‌های رگرسیونی با انجام شبیه سازی خواهیم پرداخت و در نهایت به تحلیل و بررسی یک مجموعه داده واقعی توسط این روش‌های جدید می‌پردازیم.

فهرست مطالب

فهرست علائم اختصاری

مجموع توان دوم باقی‌مانده‌ها ..... RSS

میانگین توان دوم خطا ..... MSE

کمترین توان دوم عادی ..... OLS

کمترین توان دوم تاوانیده ..... PLS

اعتبارسنجی متقابل CV

اعتبارسنجی متقابل با حذف یک داده ..... $CV_{(n)}$

اعتبارسنجی متقابل k لایه ..... $CV_{(k)}$

فاصله استوار ..... $RD(x_i)$

کارایی نسبی مجانبی ..... ARE

مجموع توان دوم کل ..... TSS

کمترین توان‌های دوم پیراسته‌ی چندگانه ..... $MC(\mu)$

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمه و تاریخچه

تحلیل رگرسیون روش آماری برای بررسی و مدل سازی ارتباط بین متغیرهاست. کاربردهای رگرسیون متعدد است و تقریبا در هر زمینه ای از جمله مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی، بیولوژی و علوم اجتماعی بکار می‌رود. در حقیقت تحلیل رگرسیونی ممکن است روش آماری با بیشترین و وسیع‌ترین کاربرد بین روش‌های آماری باشد. مدل‌های رگرسیونی برای اهدافی مانند :

۱. توصیف داده‌ها

۲. برآورد پارامترها

۳. پیشگویی و برآورد

مورد استفاده قرار می‌گیرند.

دو بخش اصلی رگرسیون در آمار وجود دارد: پارامتری و ناپارامتری. یکی از مسائلی که در بحث استنباط آماری مطرح می‌شود برآورد پارامترها و تشخیص مدل صحیح با انتخاب متغیر در حالت آمار پارامتری است. انتخاب متغیر، یکی از مهم‌ترین موضوعات در مدل‌سازی آماری محسوب می‌شود که کاربرد وسیعی در تحلیل‌های آماری دارد. بسیای از روش‌های انتخاب متغیر با استفاده از مدل رگرسیون خطی انجام می‌شود. اما کاملا آشکار است که روش کمترین توان‌های دوم اغلب در تفسیر و پیش‌بینی مدل ناتوان عمل می‌کند. روش‌هایی برای حل این مشکل ارائه شده که در بین آنها مدل رگرسیون تاوانیده از اهمیت خاصی برخوردار است.

یکی از این روش‌های تاوانیده، رگرسیون تاوانیده‌ی ستیغی است که برای اولین بار در سال ۱۹۶۲ توسط هورل [۱] معرفی شد که عملکرد آن انقباض پارامترها به سمت صفر است. در حالت‌هایی که تعداد زیادی متغیر در مدل وجود دارد روش کمترین توان‌های دوم کارایی خود را از دست می‌دهد و تفسیر مدل دشوار خواهد شد. برای رفع این مشکل در سال ۱۹۹۶ تیبشیرانی [۲] رگرسیون تاوانیده لاسو را معرفی کرد. یکی از ویژگی‌های لاسو صفر کردن بعضی از پارامترها و یا به عبارتی انتخاب متغیر می‌باشد.

یکی دیگر از مشکلاتی که برآوردگر کمترین توان‌های دوم با آن مواجه است، حساسیت بیش از اندازه آن نسبت به نقاط دورافتاده موجود در مجموعه‌ی مشاهدات می‌باشد. یک روش برای حل آن رگرسیون استوار است که توسط هیوبر [۳] در سال ۱۹۸۱ ارائه شد و بعد از او هامپل و همکاران [۴] در سال ۱۹۸۶ ارزشمندترین منبع آمار استوار را ارائه دادند.

رگرسیون استوار برای کاهش اثر مشاهداتی به کار می‌رود که اگر روش کمترین توان‌های دوم بکار گرفته شود تاثیرگذاری بالایی خواهد داشت. یک رگرسیون استوار علاوه بر حساس نبودن نسبت به نقاط دورافتاده وقتی که مشاهدات دارای توزیع نرمال هستند کارایی ۹۰ تا ۹۵ درصد نسبت به روش برآورد کمترین توان‌های دوم خواهد داشت. روش‌های استوار به مشاهدات وزن نابرابر اختصاص می‌دهند و به طور کلی مشاهداتی که باقیمانده‌های بزرگ را تولید می‌کنند، بوسیله‌ی این روش کم وزن‌تر می‌شوند.

یک روش استوار "رگرسیون با انقباض نقاط دورافتاده" (ROS) [۵] نامیده می‌شود که توسط شیفنگ [۶] و روشن [۷] در سال ۲۰۱۳ معرفی شده است. این روش باعث بهبود روش‌های استوار دیگر مانند LTS [۸] می‌شود، به این دلیل که برآوردگر ROS به دو ویژگی ماکزیمم مقدار فروریزش و کارایی کامل مجانبی بطور هم‌زمان دست می‌یابد. بعلاوه پیچیدگی محاسباتی آن

---

[۱] Horel

[۲] Tibshirani

[۳] Huber

[۴] Hampel et al

[۵] Regression with Outlier Shrinkage

[۶] Shifeng

[۷] Roshan

[۸] Least Trimmed Squares

کمتر از LTS می‌باشد. آنها همچنین این روش را برای برآوردهای تنک تعمیم دادند بطوری که هم‌زمان انتخاب متغیر و برآورد پارامترها را انجام می‌دهد. آنها این برآوردگر را "رگرسیون تنک با انقباض نقاط دورافتاده" [۹] نامیدند که به کارایی کامل مجانبی و برآورد ضرایب غیر صفر دست می‌یابد.

[۹] Sparse Regression with Outlier Shrinkage

فصل ۲

رگرسیون خطی و انتخاب متغیر

۲-۱ مقدمه

رگرسیون در لغت به معنی بازگشت و در اصطلاح آماری به معنای یافتن رابطه بین متغیرهای مستقل و وابسته می‌باشد. رگرسیون به دو دسته تقسیم می‌شود: یا هدف یافتن رابطه یک متغیر خاص با متغیر وابسته است که به آن رگرسیون ساده گفته می‌شود و یا هدف یافتن رابطه بین چند متغیر مستقل و متغیر وابسته است که به این شرایط رگرسیون چندگانه می‌گویند. در آمار پارامتری رابطه بین متغیرهای مستقل و وابسته از طریق تابع پارامتری f بیان می‌شود. اما در رگرسیون ناپارامتری رابطه بین متغیرهای مستقل و وابسته ازطریق یک تابع ناپارامتری و نامعلوم مانند $m(.)$ به صورت $y_i = m(\mathbf{x}_i) + \epsilon_i$ نشان داده می‌شود. در بعضی از مواقع که با مسائلی همچون بعد بالا و یا هم‌خطی مواجه می‌شویم استفاده از رگرسیون خطی ناکارامد بوده و برآورد حاصل از این روش مناسب نخواهد بود. در چنین شرایطی استفاده از روش جدیدی با عنوان رگرسیون تاوانیده [۱] مطرح خواهد شد که در ادامه به معرفی آن خواهیم پرداخت.

فرض می‌کنیم $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ متغیرهای مستقل و y_i؛ $1 \leq i \leq n$ ، متغیر وابسته بوده و $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$ خطای تصادفی باشد. اگر زوج مرتب این مشاهدات را به صورت $(\mathbf{x}_i, y_i)$ $i = 1, ..., n$ در نظر بگیریم آنگاه می‌توان رابطه بین هر زوج را بصورت $y_i = f(\mathbf{x}_i) + \epsilon_i$

[۱] Penalized regression

نشان داد. در رگرسیون پارامتری هدف برآورد تابع $f(\mathbf{x}_i;\boldsymbol{\beta})$ است. برای برآورد $f(\mathbf{x}_i;\boldsymbol{\beta})$ لازم است تابع هدف زیر را نسبت به $f(\mathbf{x}_i;\boldsymbol{\beta})$ مینیمم کنیم

$$Q(\boldsymbol{\beta}) = \sum_{i=۱}^{n} L(y_i - f(\mathbf{x}_i;\boldsymbol{\beta})). \qquad (۱-۲)$$

در رابطه (۱-۲) تابع $L(.)$ را تابع زیان می‌نامند که دارای سه حالت خاص

۱. کمترین توان دوم خطا: $L(\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}^۲$

۲. کمترین قدرمطلق خطا: $L(\boldsymbol{\epsilon}) = \mid \boldsymbol{\epsilon} \mid$

۳. کمترین توان دوم/قدرمطلق خطای موزون: $L(\boldsymbol{\epsilon}) = \mathbf{W}\rho(\boldsymbol{\epsilon})$

می‌باشد.

۲-۲ رگرسیون خطی

اگر تابع $f(\mathbf{x}_i,\boldsymbol{\beta})$ دارای رابطه خطی باشد با جایگذاری آن در y_i خواهیم داشت:

$$y_i = \sum_{j=۰}^{p} \beta_j \mathbf{x}_{ij} + \epsilon_i$$

در این حالت با برآورد $f(\mathbf{x}_i,\boldsymbol{\beta})$ پارامترهای مجهول برآورد خواهند شد. یکی از معروف‌ترین روش‌های برآورد پارامترها، روش کمترین توان‌های دوم عادی [۲] است که در زیر به توضیح مختصری درباره آن می‌پردازیم.

۲-۲-۱ روش کمترین توان‌های دوم عادی

کمترین توان‌های دوم عادی (OLS) یک روش برای برازش مدل به داده‌های مشاهده شده است که برای این مقصود از مینیمم کردن مجموع توان دوم باقیمانده‌ها استفاده می‌کند. تابع

[۲] Ordinary Least Square