

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشکده علوم
گروه ریاضی

پایان نامه کارشناسی ارشد رشته آمار گرایش آمار ریاضی

رگرسیون با انقباض نقاط دورافتاده

استاد راهنما

دکتر حیدرعلی مردانی فرد

پژوهشگر

شهره شریفی میناب

بهمن ماه ۱۳۹۳

حمایت از حقوق پدیدآوردگان

پایان نامه حاضر، حاصل پژوهشهای نگارنده در دوره کارشناسی ارشد رشته آمار گرایش آمار ریاضی است که در بهمن ماه ۱۳۹۳ در دانشکده علوم دانشگاه یاسوج به راهنمایی دکتر حیدرعلی مردانی فرد و مشاوره دکتر آرش اردلان 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 ریاضی

رگرسیون با انقباض نقاط دورافتاده

استاد راهنما

دکتر حیدرعلی مردانی فرد

پژوهشگر

شهره شریفی میناب

بهمن ماه ۱۳۹۳



رگرسیون با انقباض نقاط دورافتاده

به وسیله

شهره شریفی میناب

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 حیدرعلی مردانی فرد | با مرتبه علمی استادیار | امضاء |
| ۲- استاد مشاور: | دکتر آرش اردلان | با مرتبه علمی استادیار | امضاء |
| ۳- استاد داور داخل گروه: | دکتر روح الله روزگار | با مرتبه علمی استادیار | امضاء |
| ۴- استاد داور خارج گروه: | دکتر علیرضا نعمت الهی | با مرتبه علمی استاد | امضاء |
| ۵- نماینده تحصیلات تکمیلی دانشگاه: | دکتر حمیدرضا رجبی | با مرتبه 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 بر ما 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 یاری‌گرشان این معنا قطعاً محقق نمی‌گشت.

در این مجال بر خود واجب و لازم می‌دانم از استاد با کمالات و شایسته، جناب آقای دکتر حیدرعلی مردانی فرد 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 توان‌نیده است که تاثیر آن بر اندازه پارامترها و میزان تمایل آنها به صفر می‌باشد.

در این راستا با ترکیب هر دو روش استوار و توان‌نیده قادر به ارائه روش جدیدی با نام "رگرسیون با انقباض نقاط دورافتاده" هستیم که هم نسبت به روش‌های دیگر از استواری بیشتری برخوردار است و هم 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 شبیه‌سازی خواهیم پرداخت و در نهایت به تحلیل و بررسی یک مجموعه داده واقعی توسط این روش‌های جدید می‌پردازیم.

فهرست مطالب

iv	فهرست علائم اختصاری
v	فهرست تصاویر
vii	فهرست جداول
۱	فصل ۱: مقدمه و تاریخچه
۴	فصل ۲: رگرسیون خطی و انتخاب متغیر
۴	۱-۲ مقدمه
۵	۲-۲ رگرسیون خطی
۵	۱-۲-۲ روش کمترین توان‌های دوم عادی
۱۴	۳-۲ رگرسیون تاوانیده
۱۵	۱-۳-۲ رگرسیون خطی تاوانیده
۱۸	۲-۳-۲ انواع رگرسیون خطی تاوانیده
۳۰	۳-۳-۲ انتخاب پارامتر تنظیم کننده
۳۳	۴-۲ انتخاب متغیر
۴۱	فصل ۳: روش‌های استوار در تحلیل رگرسیون خطی
۴۱	۱-۳ مقدمه
۴۲	۲-۳ روش‌های رگرسیون استوار
۴۴	۱-۲-۳ برآوردگر کمترین توان‌های دوم پیراسته
۴۵	۲-۲-۳ برآوردگر کمترین قدرمطلق انحرافات
۴۷	۳-۲-۳ برآوردگر کمترین قدرمطلق انحرافات موزون

۵۰	۴-۲-۳ برآوردگر کمترین توان‌های دوم موزون استوار و کارا
۵۴	۵-۲-۳ برآوردگر آستانه θ براساس فرایند تکرار برای تشخیص نقطه دورافتاده
۶۰	۶-۲-۳ برآوردگر ROS
۶۱	۷-۲-۳ مدل تغییر جای میانگین
۶۳	۸-۲-۳ مقدار فروریزش ROS
۶۳	۹-۲-۳ ویژگی مجانبی برآوردگر ROS
۶۴	۱۰-۲-۳ انتخاب μ
۶۴	۳-۳ شبیه سازی
۷۰	فصل ۴: روش‌های استوار رگرسیون خطی برای انتخاب متغیر
۷۰	۱-۴ مقدمه
۷۱	۲-۴ برآوردگر کمترین قدرمطلق انحرافات با تاوان لاسو
۷۳	۳-۴ برآوردگر کمترین قدرمطلق انحرافات موزون با تاوان لاسو
۷۵	۴-۴ برآوردگر SROS
۷۶	۱-۴-۴ مقدار فروریزش بهینه برای برآورد تنک
۷۷	۲-۴-۴ میزان فروریزش SROS
۷۷	۳-۴-۴ ویژگی مجانبی SROS
۷۸	۴-۴-۴ روش محاسبه SROS
۷۹	۵-۴ برآورد پارامتر برای داده‌های شبیه‌سازی شده
۸۲	۶-۴ مثال با استفاده از داده واقعی
۸۵	۷-۴ نتیجه گیری
۸۶	پیوست آ: برنامه‌های نوشته شده برای برآورد پارامترها با استفاده از نرم‌افزار R
۸۶	آ-۱ برنامه نمودارها و جداول فصل‌های قبل
۱۳۶	پیوست ب: تعاریف و مفاهیم مورد نیاز
۱۵۷	واژه‌نامه فارسی به انگلیسی
۱۵۸	واژه‌نامه انگلیسی به فارسی

فهرست علائم اختصاری

RSS	مجموع توان دوم باقی مانده‌ها
MSE	میانگین توان دوم خطا
OLS	کمترین توان دوم عادی
PLS	کمترین توان دوم تاوانیده
CV	اعتبارسنجی متقابل
$CV_{(n)}$	اعتبارسنجی متقابل با حذف یک داده
$CV_{(k)}$	اعتبارسنجی متقابل k لایه
$RD(x_i)$	فاصله استوار
ARE	کارایی نسبی مجانبی
TSS	مجموع توان دوم کل
$MC(\mu)$	کمترین توان‌های دوم پیراسته‌ی چندگانه

فهرست تصاویر

- ۱-۲ نمودار بررسی وجود هم خطی بین دو متغیر Age و Limit ۸
- ۲-۲ نمودار بررسی وجود هم خطی بین دو متغیر Limit و Rating ۹
- ۳-۲ برازش خط رگرسیونی به داده‌های اولیه ۱۰
- ۴-۲ برازش خط رگرسیونی به داده‌ها در حضور نقطه دورافتاده ۱۱
- ۵-۲ برازش خط رگرسیونی به داده‌های اولیه ۱۲
- ۶-۲ برازش خط رگرسیونی به داده‌ها در حضور نقطه نافذ ۱۲
- ۷-۲ برازش خط رگرسیونی در حضور نقطه نافذ خوب ۱۳
- ۸-۲ برازش خط رگرسیونی ساده به داده‌ها ۱۶
- ۹-۲ نمودار تغییر ضرایب با تغییر پارامتر تنظیم کننده ستیغی ۱۹
- ۱۰-۲ فضای کانتور ستیغی ۲۲
- ۱۱-۲ نمودار تغییر ضرایب با تغییر پارامتر تنظیم کننده لاسو ۲۵
- ۱۲-۲ فضای کانتور لاسو ۲۶
- ۱۳-۲ نمودار انتخاب پارامتر تنظیم کننده ۳۳
- ۱۴-۲ آستانه نرم ۳۶
- ۱۵-۲ فضای کانتور لاسو سازوار ۳۸
- ۱۶-۲ نمودار رگرسیون لاسو سازوار ۴۰
- ۱-۳ تعداد تماس‌های بین المللی از بلژیک با برازش خط *OLS* ۴۳
- ۲-۳ نمودار برازش خط‌های رگرسیونی *OLS* و *LTS* به داده‌های starsCYG ۴۵
- ۳-۳ تعداد تماس‌های بین المللی از بلژیک با برازش خط *OLS* و *LAD* ۴۷
- ۴-۳ نمودار برازش خط *WLAD* به داده‌های صندوق بازنشستگی کشور هلند ۵۰
- ۵-۳ نمودار شاخص باقیمانده‌های مربوط به برآورد پارامترها با ۴ روش مختلف ۶۶

- ۳-۶ نمودار جعبه‌ای برای معیار SE زمانی که $\beta = (3, 2, 1/5, 1, 1, 1, 1, 1)^T$ است. ۶۹
- ۴-۱ نمودار جعبه‌ای برای معیار SE زمانی که $\beta = (3, 2, 1/5, 0, 0, 0, 0, 0)^T$ است. ۸۱
- ۴-۲ نمودار باقیمانده‌ها ۸۴
- ب-۱ نمودار انتخاب با روش بهترین زیرمجموعه ۱۴۱
- ب-۲ نمودار انتخاب با روش گام به گام رو به جلو ۱۴۳
- ب-۳ نمودار انتخاب با روش گام به گام رو به عقب ۱۴۵
- ب-۴ نمودار رگرسیون لاسو بروش *LARS* ۱۴۹

فهرست جداول

۹	۱-۲
۹	۲-۲
۲۳	۳-۲
۲۷	۴-۲
۳۵	۵-۲
۳۹	۶-۲
۶۵	۱-۳
	۱-۴
	۲-۴
	۱-ب
	۲-ب
	۳-ب

CS: نرخ انتخاب صحیح مدل، CR: نرخ کاهش متغیر صحیح، AN:

متوسط تعداد متغیرهای انتخاب شده

فصل ۱

مقدمه و تاریخچه

تحلیل رگرسیون روش آماری برای بررسی و مدل سازی ارتباط بین متغیرهاست. کاربردهای رگرسیون متعدد است و تقریباً در هر زمینه ای از جمله مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی، بیولوژی و علوم اجتماعی بکار می رود. در حقیقت تحلیل رگرسیونی ممکن است روش آماری با بیشترین و وسیع ترین کاربرد بین روش های آماری باشد. مدل های رگرسیونی برای اهدافی مانند:

۱. توصیف داده ها

۲. برآورد پارامترها

۳. پیشگویی و برآورد

مورد استفاده قرار می گیرند.

دو بخش اصلی رگرسیون در آمار وجود دارد: پارامتری و ناپارامتری. یکی از مسائلی که در بحث استنباط آماری مطرح می شود برآورد پارامترها و تشخیص مدل صحیح با انتخاب متغیر در حالت آمار پارامتری است. انتخاب متغیر، یکی از مهم ترین موضوعات در مدل سازی آماری محسوب می شود که کاربرد وسیعی در تحلیل های آماری دارد. بسیاری از روش های انتخاب متغیر با استفاده از مدل رگرسیون خطی انجام می شود. اما کاملاً آشکار است که روش کمترین توان های دوم اغلب در تفسیر و پیش بینی مدل ناتوان عمل می کند. روش هایی 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 تیشیرانی^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 مجانی بطور هم‌زمان دست می‌یابد. بعلاوه پیچیدگی محاسباتی آن

Horel^۱Tibshirani^۲Huber^۳Hampel et al^۴Regression with Outlier Shrinkage^۵Shifeng^۶Roshan^۷Least Trimmed Squares^۸

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.

فصل ۲

رگرسیون خطی و انتخاب متغیر

۱-۲ 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 از طریق یک تابع ناپارامتری و نامعلوم مانند $m(\cdot)$ به صورت $y_i = m(x_i) + \epsilon_i$ نشان داده می‌شود. در بعضی از مواقع که با مسائلی همچون بعد بالا و یا هم‌خطی مواجه می‌شویم استفاده از رگرسیون خطی ناکارآمد بوده و برآورد حاصل از این روش مناسب نخواهد بود. در چنین شرایطی استفاده از روش جدیدی با عنوان رگرسیون توانیده^۱ مطرح خواهد شد که در ادامه به معرفی آن خواهیم پرداخت.

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

نشان داد. در رگرسیون پارامتری هدف برآورد تابع $f(x_i; \beta)$ است. برای برآورد $f(x_i; \beta)$ لازم است تابع هدف زیر را نسبت به $f(x_i; \beta)$ مینیمم کنیم

$$Q(\beta) = \sum_{i=1}^n L(y_i - f(x_i; \beta)). \quad (1-2)$$

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

$$1. \text{ کمترین توان دوم خطا: } L(\epsilon) = \epsilon^2$$

$$2. \text{ کمترین قدرمطلق خطا: } L(\epsilon) = |\epsilon|$$

$$3. \text{ کمترین توان دوم/قدرمطلق خطای موزون: } L(\epsilon) = W\rho(\epsilon)$$

می‌باشد.

۲-۲ رگرسیون خطی

اگر تابع $f(x_i, \beta)$ دارای رابطه خطی باشد با جایگذاری آن در y_i خواهیم داشت:

$$y_i = \sum_{j=0}^p \beta_j x_{ij} + \epsilon_i$$

در این حالت با برآورد $f(x_i, \beta)$ 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.

۱-۲-۲ 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