بسمه تعالی

دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

پایان نامه کارشناسی ارشد آمارمحض

عنوان:

رگرسیون حداقل قدرمطلق انحرافات وزنی استوار

دانشجو:

منیژه محمودی

استاد راهنما:

جناب آقای دکتر صادق رضایی

استاد مشاور:

جناب آقای دکتر سعید رضاخواه

مهر ۱۳۸۵

بسمه تعالی

دانشگاه صنعتی امیر کبیر
(پلی تکنیک تهران)

فرم اطلاعات پایان نامه
کارشناسی ارشد و دکترا

معاونت پژوهشی

تاریخ :
پیوست :

نام و نام خانوادگی : منیزه محمودی **دانشجوی آزاد** ■ **بورسیه** ☐ **معادل** ☐

شمارهٔ دانشجوئی : ۸۳۱۱۳۱۰۹ **دانشکده :** ریاضی و علوم کامپیوتر **رشتهٔ تحصیلی :** آمارریاضی

نام و نام خانوادگی استاد راهنما : جناب آقای دکتر صادق رضایی

عنوان پایان نامه به فارسی: *رگرسیون حداقل قدرمطلق انحرافات وزنی استوار*

عنوان پایان نامه به انگلیسی: *Robust weighted least absolute deviations regression*

نوع پروژه : **کارشناسی ارشد :** ■ **کاربردی** ■ **بنیادی** ☐ **توسعه ای** ☐ **نظری** ☐
دکتـــــری ☐

تاریخ شروع : ۸۴/۸/۲۰ **تاریخ خاتمه :** 5 /۸۵/7 **تعداد واحد :** ۶

سازمان تأمین کننده اعتبار : ندارد

واژه های کلیدی به فارسی : رگرسیون استوار؛ *رگرسیون حداقل قدرمطلق انحرافات؛ رگزسیون جندک*

واژه های کلیدی به انگلیسی : Robust regresion , Least absolute deviations regression, Quantile regression

نظرها و پیشنهادها به منظور بهبود فعالیتهای پژوهشی دانشگاه : دسترسی به منابع علمی ازجمله مقالات ومجلات لاتین معتبر آسانتر شود .

استاد راهنما : جناب آقای دکتر صادق رضایی

دانشجو : منیزه محمودی

امضاء استاد راهنما : **تاریخ :**

نسخهٔ ۱ : معاونت پژوهشی
نسخهٔ ۲: کتابخانه و به انضمام دوجلد پایان نامه به منظور تسویه حساب با کتابخانه و مرکز اسناد و مدارک علمی

تقدیم به پدر و مادر مهربان و خواهر عزیزم

که سرچشمه های امید و مهربانی در زندگیم هستند و تشکر از

زحمات و مهر بی دریغ این فرشتگان که مرا در تمام مراحل زندگیم

یاری کردند.

تشکر و قدردانی

فرصت مناسبی است برای سپاسگزاری و قدردانی از زحمات ارزنده اساتید محترم، جناب آقای دکتر صادق رضایی و جناب آقای دکتر سعید رضاخواه که با راهنماییهای مفیدشان، همچون پدری دلسوز و مهربان مرا در اتمام این پایان نامه یاری رساندند. همچنین سپاس صمیمانه خود را به جناب آقای دکتر اسماعیل خرم و جناب آقای دکترحمید پزشک که زحمت داوری این پایان نامه را برعهده گرفته اند، تقدیم می کنم.

فهرست مطالب:

فصل سوم رگرسیون استوار وانواع برآوردگرهای آن

فصل چهارم رگرسیون حداقل قدرمطلق انحرافات وزنی

پیوستها

چکیده :

مقاله اصلی که در این پایان نامه مبنای کار در نظر گرفته شده است تحت عنوان

" Robust Weighted LAD Regression "

است که توسط Giloni، Sengupta، Simonoff در سال ۲۰۰۶ ارائه شده است.

زمانی که در مدل رگرسیون خطی نقاط پرت و دورافتاده وجود داشته باشد، یا مشاهدات از توزیع غیر نرمال تبعیت کنند؛ شیوه حداقل مربعات، دیگر شیوه خوبی برای برآورد پارامترها نیست؛ زیرا این برآوردگر نسبت به مشاهدات غیرمعمول بسیار حساس است. بنابراین شیوه رگرسیون استوار با تعداد زیادی برآوردگر پیشنهاد شده است.

یکی از قدیمی ترین پیشنهادات، شیوه حداقل قدرمطلق انحرافات (LAD) بوده است، که ضرایب رگرسیونی در آن از طریق مینیمم کردن مجموع قدرمطلق باقیمانده ها برآورد می شوند.

البته از رگرسیون حداقل قدرمطلق انحرافات به عنوان یک جایگزین قوی برای شیوه حداقل مربعات چشم پوشی شده است، به این دلیل که این برآوردگر به شدت می تواند بوسیله یک تک مشاهده تحت تأثیر قرار گرفته شود، زیرا این برآوردگر دارای نقطه فروریزش برابر ۱/n است. (n حجم نمونه است.)

هدف اصلی در این پایان نامه این است که نشان دهیم، با انتخاب وزنهای منطقی می توان برآوردگر حداقل قدرمطلق انحرافات وزنی را که دارای نقطه فروریزش بالاتری نسبت به برآوردگرLAD است، را بدست آورد. سپس ویژگیهای این برآوردگر را مورد بررسی قرار داده و همچنین کاربرد این برآوردگر استوار را روی داده های واقعی نشان می دهیم.

لازم به ذکر است که ما مقاله ای تحت عنوان " رگرسیون حداقل قدرمطلق انحرافات وزنی استوار" را به مجله علمی- پژوهشی دانشگاه الزهرا (س) ارسال کرده ومنتظر داوری نهایی آن می باشیم، همچنین مقاله ای را تحت همان عنوان در هشتمین کنفرانس آمار ایران در شیراز در تاریخ ۸۵/۵/۲ ارائه دادیم.

برای اولین بار هم مقاله ای را تحت عنوان

" Statistical Inferences on Robust Weighted LAD Regression "

در مورد استنباط آماری رگرسیون حداقل قدرمطلق انحرافات وزنی استوار به یک ژورنال معتبر Statistica Sinica ارسال کرده و منتظر داوری آنها می باشیم.

علا ئم اختصاری

IRLS: Iteratively Reweighted Least Squares

LAD: Least Absolute Deviation

LMS: Least Median Squares

LTS: Least Trimmed Squares

M-stimator: Maximum likelihood estimator

MAD: median absolute deviations

OLS: Ordinary Least Squares

WLAD: Weighted Least Absolute Deviations

فصل ۱

کلیات

۱-۱ مقدمه

تحلیل رگرسیونی یکی از فنون آماری است که بیشترین کاربرد را در علوم اجتماعی، پزشکی، اقتصادی، کشاورزی و در بسیاری از زمینه های دیگر علمی و علوم کاربردی داراست. تحلیل رگرسیونی شامل مفاهیمی می شود که به سادگی قابل درک هستند و عملاً در اکثر نرم افزارهای آماری قابل اجرا می باشند و برای بدست آوردن ارتباط بین متغیرها کاربرد فراوانی دارد.

در اغلب برازشهای رگرسیونی از شیوه حداقل مربعات معمولی (OLS [۱]) استفاده می شود، اما زمانی که خطاها دارای توزیع نرمال نباشند و یا مجموعه داده ها شامل داده های پرت باشند؛ روش حداقل مربعات معمولی، دیگر کارا نیست زیرا این شیوه به داده های پرت حساس می باشد و بایستی از روشهای استوار برای برآورد پارامترها استفاده نمود. رگرسیون استوار معمولاً به روشی گفته می شود که نه تنها وقتی خطاها دارای توزیع نرمال است و مشاهدات پرت در مدل حضور ندارند، خوب عمل می کند بلکه نسبت به انحرافات کوچک از فرض نرمال بودن و نسبت به حضور نقاط پرت در مدل نیز حساس نمی باشد. رگرسیون استوار دارای تعداد زیادی برآوردگر است و تکنیکهای آنها مکمل تکنیک کمترین مربعات می باشند به طوریکه وقتی توزیع خطاها نرمال است و مشاهدات پرت در مدل حضور نداشته باشند، جوابهای آنها مشابه با جواب رگرسیون حداقل مربعات است.

یکی از این شیوه ها، شیوه حداقل قدر مطلق انحرافات (LAD [۲]) است که ضرایب رگرسیونی را از

۱- **Ordinary Least Squares**

۲- **Least Absolute Deviations**

از طریق مینیمم کردن مجموع قدرمطلق باقیمانده ها برآورد می کند. این روش تقریباً ۵۰ سال قبل از روش حداقل مربعات، در سال ۱۷۵۷ توسط Boscovich معرفی شد. پس از آن Laplace آنرا ۳۰ سال بعد پذیرفت ولی خیلی زود روش حداقل مربعات به دلیل سادگی محاسبات، برآن سایه افکند. اما امروزه که برای محاسبات محدودیتی وجود ندارد و از طرف دیگر به دلیل حساسیت رگرسیون OLS[۱] نسبت به حضور نقاط پرت و فرض غیر نرمال خطاها و استواری رگرسیون LAD[۲] در این موارد، موجب شده است که بار دیگر این برآوردگر در زمره برآوردگرهای خوب واستوار قرار بگیرد؛ ولی به این دلیل که این شیوه نسبت به سایر برآوردگرهای استوار، دارای نقطه فروریزش کمتری است، (نقطه فروریزش یک ویژگی مهم برآوردگرهای استوار است که به عنوان ملاکی برای مقایسه آنها استفاده می شود)، از کاربرد آن کاسته است.

به همین دلیل در صدد برآمدیم تا در جهت بالابردن استواری برآوردگر LAD این تحقیق را انجام دهیم. در این پایان نامه، شیوه جدید رگرسیون استوار را به نام رگرسیون حداقل قدرمطلق انحرافات وزنی (WLAD[۳]) را ارائه می دهیم که این شیوه دارای نقطه فروریزش بالاتری نسبت به رگرسیون LAD است و نشان می دهیم که با انتخاب وزنهای ساده و منطقی می توانیم استواری آنرا افزایش دهیم.

۱- **Ordinary Least Squares**

۲- **Least Absolute Deviations**

۳- **Weighted Least Absolute Deviations**

مقاله اصلی که در اینجا مبنای کار در نظر گرفته شده است تحت عنوان،

" Robust Weighted Least Absolute Deviations Regression "

است که توسط Giloni، Simonoff و Sengupta در سال ۲۰۰۶ برای اولین بار در این زمینه ارائه شده است.

از طرف دیگر چون استنباط آماری در مباحث رگرسیونی دارای اهمیت فراوانی است و تاکنون هیچ مقاله ای درباره استنباط آماری رگرسیون WLAD ارئه نشده است، طبق این ضرورت، ما برای اولین بار مطالعات و تحقیقاتی را در این زمینه انجام داده و مقاله ای را با عنوان

" Statistical Inference on Robust Weighted LAD Regression "

ارائه و به ژورنال معتبر علمی ارسال کردیم.

۱-۲ تاریخچه

این بخش را به مروری بر تاریخچه مقالاتی که در این پایان نامه بکاربردیم، اختصاص می دهیم.

Huber در سال ۱۹۷۳ در مقاله ای تحت عنوان " Robust Regression " و در سال ۱۹۸۱ در مقاله " Robust Statistics " ، Rousseeuw در سال ۱۹۸۱ در کتابی تحت عنوان:

" Robust Regression and Outlier Detection "

و Chatterjee و Hadi در سال ۱۹۸۸ در کتاب

" Sensitivity analysis in linear regression "

تحقیقات وسیعی را در زمینه رگرسیون استوار و ضرورت استفاده از آن ارائه کرده اند.

برای تکمیل تحقیقات رگرسیون WLAD[۱] به تاریخچه رگرسیون حداقل قدرمطلق انحرافات نیازمندیم. رگرسیون LAD[۲] دارای قدمت بسیار زیادی است، Dielman در سال ۲۰۰۵ در مقاله ای تحت عنوان " Least absolute value regression " یک مروری بر تحقیقات گسترده ای که تا آن زمان در زمینه رگرسیون LAD انجام شده بود را ارائه کرده است. او به تاریخچه کامل رگرسیون LAD ، شیوه های پیشنهادی برای برآورد آن و به محاسبات آنها اشاره کرده است.

Boscovich در سال ۱۷۵۷ برای اولین بار این شیوه را پیشنهاد کرد، در سالهای اخیر الگوریتمهای بسیار زیادی برای برآورد LAD پیشنهاد شده است که یکی از آنها، الگوریتم سیمپلکس بر اساس

۱- **Weighted Least Absolute Deviations**

۲- **Least Absolute Deviations**

مسئله ثانویه است که توسط Barradale و Roberts در سال ۱۹۷٤ بیان شد و Dielman در سال ۱۹۹۲ آنرا اصلاح کرد. Bloomfield و Steiger در سال ۱۹۸۰ الگوریتم دیگری را به نام الگوریتم حداقل مربعات موزون تکراری (IRLS) بیان کردند که Rousseeuw وLeroy در سال ۱۹۸۷، Portnoy و Koenker در سال ۱۹۹۷، Olive و Hawkins در سال ۲۰۰۲، Koenker در سال ۲۰۰٤ به تصحیح شیوه های ذکر شده پرداختند.

همانطور که در مقدمه ذکرشد؛ مقدار فروریزش، یک ویژگی مهم برآوردگر استوار است که Donoho و Huber (۱۹۸۳)، He et al (۱۹۹۰)، Mizera و Muller (۲۰۰۱) شیوه هایی را برای محاسبه مقدار فروریزش رگرسیون LAD پیشنهاد کردند که Giloni و Padberg در سال ۲۰۰٤ در مقاله ای با عنوان

" The finite sample breakdown point of L_1 regression "

شیوه ساده تری را برای مقدار فروریزش رگرسیون LAD در یک نمونه متناهی ارئه کردند.

برای افزایش مقدار فروریزش این برآوردگر، Rousseeuw و Zomren (۱۹۹۲) در مقاله ای با نام " A comparison of some quick algorithms " و همچنین Huber و Rousseeuw در سال ۱۹۹۷ در مقاله

" Robust regression with both continues and binary regressors "

طرحهای وزنی را پیشنهاد کردند که اولین طرح به دلیل دقت کم آن و دومی به دلیل محاسبات زیاد آن موجب شدند که Giloni، Simonoff، Sengupta (۲۰۰٦) یک طرح وزنی منطقی را پیشنهاد کنند که باعث افزایش مقدار فروریزش آن می شود و محاسبات ساده تری هم دارد و آنها این برآوردگر جدید را برآوردگر استوار WLAD معرفی کردند.

لازم به ذکر است که با استفاده از مقاله Dielman و Pfaffenberger (۱۹۹۲) با عنوان

"A further comparison of tests of hypotheses in LAV regression "

و مقاله Dielman (۱۹۹٦) تحت عنوان

" A note on hypothesis testing in LAV [1] multiple regression "

ومقاله Dielman (۲۰۰۵) با عنوان " Least absolute value regression " که آزمون فرضهایی را برای رگرسیون LAD پیشنهاد کردند و با استفاده ازمقاله Koenker و Bassett (۱۹۷۸) با عنوان" Regression Quantile " و مقاله Zhou وPortnoy (۱۹۹۸) تحت عنوان " Statistical inferences on regression quantile " ، که فاصله اطمینانی را برمبنای رگرسیون چندک برای رگرسیون LAD ارائه دادند، توانستیم برای اولین بارفاصله اطمینان و آزمون فرضی را برای رگرسیون WLAD تحت مقاله ای جدید پیشنهاد کنیم.

۱- Least Absolute Values

۱-۳ ساختار

ساختار پایان نامه به صورت زیر است:

فصل دوم به تعریف مشاهدات مؤثر و انواع آنها ، شیوه های شناخت مشاهدات مؤثر همراه با مثال اختصاص دارد.

در فصل سوم به رگرسیون استوار، کاربرد آن ومعرفی تعدادی ازبرآوردگرهای استوار M ، LMS [۱] و LTS [۲] می پردازیم و برنامه های نرم افزاری Splus آنها را در مثالهایی ارائه می دهیم.

فصل چهارم به معرفی رگرسیون حداقل قدر مطلق انحرافات، ویژگیهای مجانبی برآوردگرآن، تعریف نقطه فروریزش آن و همچنین به معرفی رگرسیون حداقل قدرمطلق انحرافات وزنی، چگونگی انتخاب وزنهای آن، چگونگی محاسبه نقطه فروریزش و ویژگیهای مجانبی آن اختصاص دارد. کاربرد مطالب این فصل را هم روی یک مثال نشان می دهیم. بخش آخر از این فصل را هم به اساس مقاله ای که برای اولین بار، ما در زمینه فاصله اطمینان و آزمون فرض برای ضرایب رگرسیون WLAD ارئه دادیم، اختصاص می دهیم.

در پایان هر فصل هم نتیجه مربوط به همان فصل را بیان می کنیم.

---

۱) **Least Median Squares**

۲) **Least Trimmed Squares**

فصل دوم

مشاهدات مؤثر