دانشکده علوم

پایان نامه کارشناسی ارشد در رشته آمار ریاضی

چند روش برآوردیابی استوار درمدل های رگرسیونی چندکی با داده های پرت

به وسیله‌ی

الهام یوسفی

استاد راهنما

دکتر امین قلم فرسای مستوفی

شهریور ۹۲

به نام خدا

اظهار نامه

اینجانب الهام یوسفی (۹۰۰۳۷۵) دانشجوی رشته‌ی آمار گرایش آمار ریاضی دانشکده‌ی علوم، اظهار می‌کنم که این پایان نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده ام نشانی دقیق و مشخصات کامل آن را نوشته ام. هم چنین اظهار می‌کنم که تحقیق و موضوع پایان نامه ام تکراری نیست و تعهد می‌نمایم که بدون مجوز از دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر قرار ندهم. کلیه حقوق این اثر مطابق با آیین نامه مالکیت فکری و معنوی متعلق به دانشگاه شیراز است.

نام و نام خانوادگی: الهام یوسفی

تاریخ و امضاء: ۱۳۹۲/۶/۱۹

به نام خدا

چند روش برآوردیابی استوار درمدل های رگرسیونی چندکی با داده های پرت

به وسیله‌ی

الهام یوسفی

پایان نامه

ارائه شده به تحصیلات تکمیلی دانشگاه شیراز به عنوان به عنوان بخشی از فعالیت-های لازم برای اخذ درجه‌ی کارشناسی ارشد

در رشته‌ی:

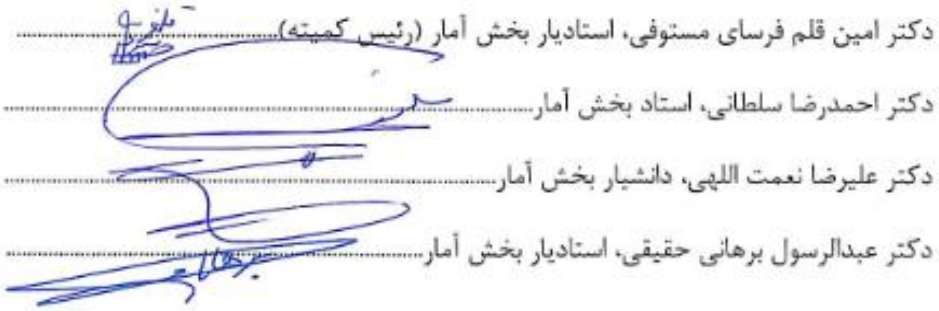
آمار ریاضی

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی کمیته‌ی پایان نامه، با درجه : عالی



شهریور ۱۳۹۲

تقدیم به دو مهربان زندگیم؛

به پدر عزیزم که عظمت هستی را در وجودش یافتم

به مادر مهربان و صبورم

که از هیچ کوششی برای سعادت من دریغ نکردند

سپاس گزاری

نخست، سپاس بی‌کران پروردگار یکتا را که هستیمان بخشید... بعد از حمد و سپاس به درگاه خداوند منان از نعمت‌های بیکرانی که بر ما ارزانی داشته است، از پدر و مادر صبور و مهربانم به خاطر تمام حمایت‌ها، تلاش‌ها و فداکاری‌هایشان که در تمام مراحل زندگی همراه من بوده اند سپاسگزارم. سپس بر خود لازم می‌دانم از زحمات استاد گرانقدرم جناب آقای دکتر قلم فرسا که پیوسته با راهنمایی‌هایشان راه گشای مشکلاتم در این مسیر بوده و از هیچ کوششی فروگذار ننمودند کمال تشکر و سپاس را داشته باشم. هم چنین از زحمات جناب آقای دکتر نعمت اللهی که با کمال صبر و حوصله در قسمت های مختلف پایان نامه در راه گشایی مشکلاتم کمک شایانی به من کرده و نکات ارزنده ای را به من آموختند، بی نهایت سپاس گزارم. از اساتید محترم جناب آقایان دکتر برهانی و دکتر سلطانی که با کمال دقت، زحمت مطالعه‌ی این پایان نامه را تقبل کرده و نکات ارزنده ای را به من گوشزد کردند، کمال تشکر و قدردانی را دارم. در انتها از امیدواری ها، همیاری ها و مهربانی های دوست همیشه همراهم... صحرا فرامرزی که پیوسته و همه جا همراهم بوده بسیار سپاس گزارم و امیدوارم روزی بتوانم قسمتی از مهربانیهایش را جبران کنم.

چکیده

چند روش برآوردیابی استوار درمدل های رگرسیونی چندکی با داده های پرت

به کوشش

الهام یوسفی

در آمار و روش‌های آماری در اقتصاد، اغلب با مجموعه داده‌هایی روبرو هستیم که شامل نقاط پرت هستند. رگرسیون چندکی خطی نیز به عنوان روشی پرکاربرد در آمار، نسبت به مشاهدات پرت، مخصوصا مشاهدات پرت موجود در متغیرهای مستقل، حساس است. در این پایان نامه ابتدا چند برآوردگر رگرسیونی را معرفی کرده و نشان می‌دهیم که چگونه می‌توان با استفاده از نقطه شکستگی استواری آنها را ارزیابی کرد. سپس درباره‌ی برآوردگرهای رگرسیون چندکی بحث می‌کنیم. به دلیل اینکه چنین برآوردگری در حضور نقاط پرت استوار نیست، برآوردگر استواری را که توسط سیزک و همکارانش در سال ۲۰۱۲ معرفی شد، مورد بررسی قرار می‌دهیم. در نهایت عملکرد رگرسیون چندکی را با رگرسیون چندکی حداقل پیراسته از طریق یک مثال عددی با هم مقایسه می‌کنیم.

واژه‌های کلیدی: برآوردگر استوار، نقطه شکستگی، رگرسیون چندکی، پیراستگی، نقطه اهرمی

نام : الهام **نام خانوادگی: یوسفی**
مقطع تحصیلی: کارشناسی ارشد **رشته و گرایش: آمار ریاضی**
استاد راهنما: دکتر امین قلم فرسا **تاریخ دفاع: ۱۳۹۲/۶/۱۹**

چکیده

چند روش برآوردیابی استوار درمدل های رگرسیونی چندکی با داده های پرت

در آمار و روش‌های آماری در اقتصاد، اغلب با مجموعه داده‌هایی روبرو هستیم که شامل نقاط پرت هستند. رگرسیون چندکی خطی نیز به عنوان روشی پرکاربرد در آمار، نسبت به مشاهدات پرت، مخصوصا مشاهدات پرت موجود در متغیرهای مستقل، حساس است. در این پایان نامه ابتدا چند برآوردگر رگرسیونی را معرفی کرده و نشان می‌دهیم که چگونه می‌توان با استفاده از نقطه شکستگی استواری آنها را ارزیابی کرد. سپس درباره‌ی برآوردگرهای رگرسیون چندکی بحث می‌کنیم. به دلیل اینکه چنین برآوردگری در حضور نقاط پرت استوار نیست، برآوردگر استواری را که توسط سیزک و همکارانش در سال ۲۰۱۲ معرفی شد، مورد بررسی قرار می‌دهیم. در نهایت عملکرد رگرسیون چندکی را با رگرسیون چندکی حداقل پیراسته از طریق یک مثال عددی با هم مقایسه می‌کنیم.

فهرست مطالب

عنوان **صفحه**

فصل اول: مقدمه

فصل دوم: معرفی چند برآوردگر رگرسیونی و بررسی استواری آن ها

فصل سوم: معرفی رگرسیون چندکی

فصل چهارم: معرفی برآوردگر استوار رگرسیون چندکی حداقل

فصل پنجم: شبیه سازی و نتیجه گیری

فهرست جدول‌ها

عنوان و شماره **صفحه**

فهرست شکل‌ها

عنوان و شماره **صفحه**

عنوان و شماره **صفحه**

فصل اول

مقدمه

۱-۱- مقدمه

۱-۱-۱- اهداف کلی رگرسیون و ایده‌ی معرفی رگرسیون چندکی

یکی از اهداف مهم بررسی‌های آماری، یافتن روابطی است که به کمک آن بتوان اثر تغییرات یک یا چند متغیر را بر روی متغیرهای دیگر پیش بینی کرد. رگرسیون در واقع روشی برای استخراج روابط موجود بین متغیرها ارائه می‌دهد. رگرسیون چندکی نیز به عنوان زیر شاخه ای از رگرسیون به دنبال این هدف است که راهکار کاملی را برای تکمیل عملکرد رگرسیون میانگین ارائه دهد.

بسیاری از مصادیق آمار کاربردی را می‌توان مثلا در پیچیدگی مدل رگرسیون خطی و روش‌های برآوردیابی مرتبط با حداقل مربعات دید. در این باره Mosteller and Tukey (1977, p. 266) گفته اند:

آنچه که خط یا منحنی رگرسیونی انجام می‌دهد، بررسی متغیر پاسخ از طریق میانگین مشاهدات با توجه به مجموعه‌ی xها می‌باشد. همچنین می‌توانیم خط یا منحنی‌های رگرسیونی مختلف را با توجه به درصد نقاط توزیع‌ها محاسبه کنیم و بنابراین تصویر کامل‌تری از نقاط بدست بیاوریم. معمولا این کار به طور متداول انجام نمی‌شود و بنابراین رگرسیون تصویر ناکاملی را ارائه می‌دهد. همانطور که میانگین تصویر ناکاملی را از یک توزیع ارائه می-دهد، بنابراین منحنی رگرسیون نیز در این باره تصویر ناکاملی را از یک مجموعه توزیع‌ها ارائه می‌دهد.

هدف رگرسیون چندکی این است که راهکار کاملی را برای تکمیل عملکرد رگرسیون میانگین ارائه دهد.

چرا برآورد حداقل مربعات در مدل رگرسیونی خطی اینقدر در آمار کاربردی نفوذ کرد؟ چه چیزی آن را به ابزار قدرتمندی تبدیل کرد؟ ایده‌های اولیه درباره‌ی انجام رگرسیون به کارهای اولیه‌ی Boscovich در نیمه‌های قرن ۱۸ام تا کارهای Edgeworth در پایان قرن ۱۹ ام برمی‌گردد. در ابتدا باید گفت که یک حقیقت مهم این است که آسانی محاسبه‌ی این برآوردگرها بسیار خوشایند است. به طور قطع این از انگیزه‌های اولیه برای موفقیت این روش بود. پس اگر خطاها در مدل به طور نرمال توزیع شده باشند، روش حداقل مربعات دارای مقدار بهینه (ماکزیمم یا مینیمم) می‌شود. همانطور که Mosteller and Tukey اظهار کرده اند، به ندرت دیده می‌شود که میانگین به تنهایی برای آنالیز آماری یک تک نمونه هدف رضایت بخشی باشد. ابزارهای پراکندگی مانند چولگی، برجستگی، نمودارهای جعبه‌ای، هیستوگرام ها و روش‌های برآوردیابی چگالی همگی برای اینکه به بینش بیشتری دست یابیم به کار گرفته می-شود. آیا می‌توان چنین کاری را در رگرسیون انجام داد؟ پس می‌توان گفت که ایده‌ی ایجاد رگرسیون چندکی می‌تواند این باشد که به جای برآورد سطح میانگین شرطی (که شامل خط، صفحه و ... است) از طریق روش حداقل مربعات، سطح‌های مختلف چندک‌های شرطی را برآورد کنیم که باعث تکمیل روش قبل می‌شود.

۱-۱-۲- اهداف ایجاد روش های رگرسیونی استوار

اخیرا در آمار و روش‌های آماری در اقتصاد بیشتر به تکنیک هایی توجه شده است که با مشاهداتی سروکار دارد که غیرمعمول هستند. که این امر از توزیع‌هایی با دنباله پهن، کدگذاری اشتباه یا ناهمگنی یافته شده در مدل ناشی می‌شود و دارای اهمیت خیلی زیاد، مخصوصا در مدل‌های رگرسیونی خطی (یا غیرخطی) و سری‌های زمانی است. هدف این است که در رگرسیون، برآوردگرهایی ارائه دهیم که تحت تاثیر نقاط پرت یا انحراف از فرضیات مدل نیستند. این برآوردگرها، برآوردگرهای استوار[1] نامیده می‌شوند. برای اثبات استواری برآوردگرهای معرفی شده از معیاری به نام نقطه شکستگی[2] استفاده می‌کنیم که در ادامه

[1] Robust estimators
[2] Breakdown point

معرفی می‌شود Rousseeuw (1997), Genton and Lucas (2003), Davies and Gather (2005). اولین برآوردگر رگرسیونی استوار با نقطه شکستگی مجانبی 1/2 برآوردگر حداقل میانه‌ی مربعات بود (LMS; Rousseeuw, 1984). برآوردگرهای حداقل مربعات پیراسته (LTS; Rousseeuw, 1985) و برآوردگرهای S ، Rousseeuw and Yohai, (1984) نمی‌توانند همزمان دارای نقطه شکستگی بالا و کارایی نسبی بالا باشند Hössjer, (1992). از سوی دیگر بسیاری از روش‌های رگرسیونی استوار کلاسیکی، ظاهرا مشکل استوار بودن را، مخصوصا برای داده هایی با توزیع نرمال، با کارایی نسبی پایین جبران می‌کنند. برای بهبود بیشتر کیفیت روش های با نقطه شکستگی بالا Rousseeuw and Leroy (1987) پیشنهاد استفاده از روش حداقل وزن مربعات (LWS)[3] را دادند که در آن به مشاهداتی که بیشتر از یک نقطه‌ی قطع[4] ثابت (نقطه ای که مشاهدات بزرگتر از آن را در یک نمونه، پرت به حساب می‌آوریم) هستند، وزن صفر داده می‌شود. این کار تغییر پذیری برآورد را نسبت به برآورد استوار اولیه کم می‌کند؛ اما توزیع مجانبی آن به برآورد استوار اولیه بستگی پیدا خواهد کرد (Welch and Ronchetti, 2002). بنابراین Gervini and Yohai (2002) پیشنهاد استفاده از حداقل مربعات وزنی با استفاده از نقطه قطع وابسته به داده‌ها را دادند که منجر به حاصل شدن برآورد استوار و کارای حداقل مربعات وزنی (REWLS)[5] شد. در این حالت باز هم توزیع مجانبی REWLS بستگی به برآورد اولیه دارد و برخی از مشاهدات از برآورد حذف می‌شوند. برای برطرف شدن این مشکلات کلاسی از روش‌های برآوردیابی دو مرحله ای را ارائه می دهیم که روش حداقل وزن مربعات دو مرحله ای (2S-LWS)[6] نامیده می‌شود. در این روش به جای استفاده از وزن صفر برای مشاهدات پرت و به کارگیری روش حداقل مربعات وزنی (WLS)[7] ، از وزن های کاملا مثبت استفاده می‌شود و حداقل وزن مربعات LWS را به جای WLS به کار می‌بریم. که منجر به حاصل شدن 2S-LWS و خواص مطلوب آن می‌شود.

گاهی اوقات رگرسیون میانگین برای برازش به داده‌ها مناسب نیست و اطلاعات کاملی از نحوه‌ی پراکندگی داده‌ها ارائه نمی‌دهد. در این گونه موارد نیاز است از رگرسیون چندکی

[3] Least Weighted Squares
[4] cut off point
[5] Robust and Efficient Weighted Least Squares
[6] 2-Step Least Weighted Squares
[7] Weighted least squares

(QR)[8] به جای رگرسیون میانگین استفاده کنیم. رگرسیون چندکی توسط Koenker and Basset (1978) معرفی شد. برآوردگر رگرسیون چندکی نسبت به دم‌های چوله و اختلافات از توزیع نرمال، استوار است. به علاوه تحت شرایط خیلی کلی توزیع مجانبی بردار ضرایب برآورد شده نرمال چند گانه است که باعث ایجاد استنباط‌های استاندارد می‌شود (Koenker and Bassett, 1978). اما این برآوردگر نسبت به حضور نقاط پرت در متغیرهای مستقل حساس است (He et al, 1990). بنابراین در این گونه موارد نیاز است برآوردگرهای استوار دیگری جایگزین رگرسیون چندکی شود. بنابراین یک رویکرد پیشنهادی را برای برآوردیابی استوار در چارچوب رگرسیون چندکی بررسی می‌کنیم که رگرسیون چندکی حداقل پیراسته (LTQR)[9] نام دارد. روش کار این برآوردگر بر اساس پیراسته کردن و حذف مشاهدات است که اثر نقاط پرت در متغیرهای مستقل را کاهش می‌دهد. روش ارائه شده برآوردگر استوار مکانی میانه را که توسط Tableman (1994 a,b) و برآوردگر حداقل انحراف مطلق پیراسته (LTA)[10] که توسط Hawkins and Olive (1999) بررسی شد، گسترش می‌دهد. برآورد LTA و LTQR چندک های رگرسیونی را برای داده هایی که از تابع هدف قطع نشده‌اند، یعنی برای زیرمجموعه‌ای از داده ها، برآورد می‌کنند و باعث ایجاد خواص استواری می‌شوند.

در فصل اول این پایان نامه ابتدا ایده و انگیزه‌ی ایجاد رگرسیون چندکی که یکی از شاخه های پرکاربرد رگرسیون است را عنوان می‌کنیم و دلیل کاملتر بودن آن نسبت به رگرسیون میانگین را بحث می کنیم. سپس مفاهیم پایه برای استفاده در فصل های بعد را خواهیم آورد. در فصل دوم چند برآوردگر پرکاربرد رگرسیون میانگین را معرفی کرده و آنها را از نظر استواری مورد بررسی قرار می‌دهیم. برای بررسی استواری از معیار نقطه شکستگی استفاده می‌کنیم. در فصل سوم رگرسیون چندکی را معرفی کرده و برآوردگر پرکاربرد آن را ارائه می‌دهیم. به دلیل ضعف برآوردگر رگرسیون چندکی در حضور نقاط پرت موجود در متغیرهای مستقل، در فصل چهارم، رگرسیون چندکی حداقل پیراسته را معرفی می‌کنیم و استواری آن را از نظر معیار نقطه شکستگی مورد بررسی قرار می‌دهیم. ایده معرفی این برآوردگر بر اساس پیراسته سازی و کوتاه سازی چندک‌های رگرسیونی است. در انتها، در فصل پنجم، ابتدا کارایی نسبی برآوردگرهای معرفی شده در فصل دوم را در حالت نمونه‌های متناهی محاسبه کرده و متوسط مربع خطا (MSE) را برای آنها تحت مدل‌های توزیعی مختلف با هم مقایسه می‌کنیم. سپس

[8] Quantile Regression
[9] Least Trimmed Quantile Regression
[10] Least Trimmed Absolute deviation

عملکرد برآوردگرهای رگرسیون چندکی و رگرسیون چندکی حداقل پیراسته را هم از طریق یک مجموعه داده واقعی و هم از طریق داده های شبیه سازی شده با هم مقایسه می‌کنیم.