
بسم الله الرحمن الرحيم

دانشگاه شهید چمران اهواز

دانشکده علوم ریاضی و کامپیوتر

شماره پایان نامه :

۹۲۱۳۶۱۰

پایان نامه کارشناسی ارشد

گرایش آمار ریاضی

عنوان :

مطالعه‌ی برآورد ریج و برخی مباحث تشخیصی در مدل‌های رگرسیونی تحت محدودیت‌های خطی تصادفی

استاد راهنما:

دکتر عبدالرحمن راسخ

استاد مشاور:

دکتر بهزاد منصوری

نگارنده :

نرگس هدایت‌پور

دی ماه سال ۱۳۹۲

چکیده

نام خانوادگی :هدایت پور	نام:نرگس	شماره دانشجویی : ۹۰۱۳۶۱۱
عنوان پایان نامه : مطالعه‌ی برآورد ریج و برخی مباحث تشخیصی در مدل‌های رگرسیونی تحت محدودیت-های خطی تصادفی		
استاد راهنما: دکتر عبدالرحمن راسخ		
استاد مشاور: دکتر بهزاد منصوری		
درجه تحصیلی: کارشناسی ارشد	رشته: آمار	گرایش: آمار ریاضی
دانشگاه : شهید چمران اهواز	دانشکده: علوم ریاضی و کامپیوتر	گروه : آمار
تاریخ فارغ التحصیلی : ۱۳۹۲/۱۰/۳۰		تعداد صفحه: ۱۳۵

کلید واژه ها : هم‌خطی، برآورد ریج، برآورد ریج تحت محدودیت‌های خطی، مباحث تشخیصی، مشاهدات موثر، مشاهدات پرت.

مطالعه‌ی مباحث تشخیصی در مدل‌های مختلف رگرسیونی دارای سابقه‌ی طولانی است. گاهی مجموعه‌ی کوچکی از داده‌ها اثرات نامتناسبی را بر روی نتایج حاصل از آنالیز رگرسیونی اعمال می‌کنند. به‌طوری که برآورد پارامترها یا مقادیر پیش‌بینی شده بیشتر تحت نفوذ این مشاهدات قرار می‌گیرند. شناسایی چنین مشاهداتی از طریق روش‌ها و معیارهایی که مبتنی بر آنالیز تأثیر است امکان‌پذیر بوده که در مقوله‌ی مباحث تشخیصی تجلی پیدا می‌کند. باهدف تعیین مشاهدات مؤثر و پرت معیارهایی همچون ماتریس پیش‌بینی، باقیمانده‌های استاندارد، DFBETAS، DFFITSS، فاصله‌ی کوک و روش انتقال میانگین پیشنهاد شده است. از سوی دیگر وجود هم‌خطی در میان متغیرهای مستقل پیامدهای نامطلوبی به‌همراه دارد. این پدیده روش کمترین توان‌های دوم را به چالش کشیده و موجب عدم کارایی برآورد حاصل می‌گردد. راه‌بردهایی همچون به‌کارگیری برآورد اریب ریج، برآورد آمیخته و ترکیبی از این دو تحت عنوان برآورد ریج آمیخته باهدف اصلاح مشکل هم‌خطی پیشنهاد شده است. حضور توأم هم‌خطی و مشاهدات مؤثر در مجموعه‌ی داده‌ها امر نامعمول و غیرمنطقی نیست بلکه یک موضوع پیچیده است. در این رساله ضمن مطالعه‌ی رگرسیون ریج و رگرسیون ریج تحت محدودیت‌های خطی تصادفی اقدام به بررسی روش‌ها و معیارهای تشخیصی مذکور جهت تعیین مشاهدات تأثیرگذار و پرت در روش رگرسیون ریج می‌کنیم. در پی آن این معیارها را به روش رگرسیونی ریج تحت محدویت‌های خطی تصادفی تعمیم داده و از این طریق مشاهدات تأثیرگذار و پرت را تشخیص می‌دهیم. در نهایت به منظور شفاف سازی و توضیح نتایج حاصل شده یک مثال عددی که طی آن مجموعه داده‌های واقعی سیمان پرتلند را مورد بررسی قرار داده‌ایم ذکر می‌کنیم.

فهرست

فصل اول

مقدمه و تاریخچه‌ای از پژوهش‌های پیشین

۱-۱ مقدمه

آنالیز رگرسیونی یکی از روش‌های آماری برای تحلیل داده‌ها است. این روش برای بررسی و به مدل در آوردن ارتباط بین متغیرها به‌کار می‌رود. کاربردهای رگرسیون متعدد است و تقریباً در هر زمینه‌ای از جمله علوم مهندسی، فیزیک، اقتصاد و... کاربرد دارد. تقریباً در همه‌ی کاربردهای رگرسیون معادله رگرسیون فقط یک تقریب از رابطه‌ی درست بین متغیرها است. هدف از تحلیل رگرسیونی و به‌کارگیری مدل رگرسیونی استنباط‌هایی چون مشخص کردن اثرات نسبی متغیرهای مستقل، پیش‌بینی و یا برآورد و انتخاب یک مجموعه مناسب از متغیرها برای مدل است.

زمانی‌که بین متغیرهای مستقل وابستگی خطی وجود نداشته باشد به سادگی می‌توان استنباط‌های ذکر شده را نتیجه گرفت. ولی متأسفانه اغلب در کاربرد مدل‌های رگرسیونی بین متغیرهای مستقل وابستگی خطی وجود دارد. وجود وابستگی خطی بین متغیرهای مستقل در مدل رگرسیونی تحت عنوان پدیده‌ی هم‌خطی[1] مطرح می‌شود. در این حالت نتایجی که براساس مدل به دست خواهیم آورد ممکن است گمراه کننده باشد. وجود هم‌خطی در مدل مشکلاتی چون ناپایداری و عدم کیفیت در برآوردهای آماری را به‌همراه دارد و همچنین دلیلی بر عدم اعتبار برآورد کمترین توان-

[1] Collinearity

های دوم است. به این ترتیب برای برخورد با اثرات نامطلوب هم‌خطی راه‌های مختلفی چون جمع‌آوری داده اضافی، تخصیص مدل مجدد و استفاده از روش‌هایی متفاوت با روش کمترین توان‌های دوم برای برآورد ضرایب رگرسیونی ارائه شده است. با توجه به اینکه پدیده‌ی هم‌خطی روش کمترین توان‌های دوم را به چالش می‌کشد، بنابراین برای برآورد ضرایب رگرسیونی روش-های دیگری پیشنهاد شده است. برآورد حاصل از این روش‌ها بر خلاف روش کمترین توان‌های دوم اریب است. پس با نادیده گرفتن ضرورت نااریبی برآوردی به‌دست خواهیم آورد که واریانس آن کمتر از واریانس برآورد کمترین توان‌های دوم است. به‌کارگیری کلاس برآوردگرهای اریب یکی از با اهمیت‌ترین روش‌ها جهت برآورد ضرایب تحت شرایط حضور هم‌خطی و با هدف رفع این مشکل محسوب می‌گردد. از جمله این برآوردها می‌توان به برآورد ریج[۱]، برآورد استاین[۲] و برآورد لیو[۳] اشاره کرد.

راه‌کار دیگری که برای اصلاح اثرات ایجاد شده‌ی ناشی از پدیده‌ی هم‌خطی ارائه شده است، استفاده از اطلاعات کمکی[٤] و یا پیشین[٥] علاوه بر اطلاعات نمونه در قالب محدودیت‌های خطی[٦] بر روی ضرایب رگرسیونی است. اعمال محدودیت‌های خطی یعنی اضافه کردن این اطلاعات به مجموعه‌ی داده‌ها باعث بهبود برآورد و به دنبال آن تقلیل اثر هم‌خطی می‌گردد. برآورد ضرایب در این شرایط تحت عنوان برآورد آمیخته[۷] معرفی می‌شود. قابل ذکر است که برآورد استاین، ریج و لیو حالت خاصی از برآورد آمیخته می‌باشند.

[1] Ridge Estimator
[2] Stein Estimator
[3] Liu Estimator
[4] Auxiliary information
[5] Prior
[6] Linear Restriction
[7] Mixed Estimator

باید توجه داشت که می‌توان از ترکیب دو روش مذکور یعنی کلاس برآوردگرهای اریب همراه با اعمال محدودیت‌های خطی برای رفع مشکل هم‌خطی بهره جست. از مقایسه‌ی معیار ماتریس میانگین توان دوم خطای[1] برآورد حاصل از ادغام دو روش با دیگر برآوردها در می‌یابیم که این برآورد عمل‌کرد بهتری دارد.

از سوی دیگر مطالعه مباحث تشخیصی[2] در مدل‌های مختلف رگرسیونی سابقه‌ی طولانی دارد. بررسی روش‌هایی چون یافتن مشاهدات پرت[3] و مشاهدات تأثیرگذار[4] از جمله این روش‌ها می‌باشند. با توجه به تأثیری که مشاهدات تأثیرگذار بر نتایج حاصل از تحلیل رگرسیونی دارند؛ بنابراین باید به دقت مشاهدات را مورد بررسی قرار داده و میزان تأثیرگذاری آن‌ها را تعیین کرد. آماردانان و محققان مطالعات گسترده‌ای در این زمینه انجام داده‌اند و برای تشخیص این مشاهدات معیارها و روش‌هایی پیشنهاد کرده‌اند که می‌توان به درایه‌های قطری ماتریس پیش‌بینی[5] (آماره‌ی نفوذ)، باقیمانده‌های استاندارد، معیار $DFBETAS$، $DFFITSS$، فاصله‌ی کوک[6] و روش انتقال میانگین[7] اشاره کرد.

حضور هم‌زمان هم‌خطی و مشاهدات با اثرات ناروا یکی از مسائل مهم ولی پیچیده است و توجه بسیاری از آماردانان به این موضوع معطوف شده است. همچنین بیان شده که وجود هم‌خطی می‌تواند اثر مشاهدات ناروا را مستور کند. با توجه به جایگاه ویژه‌ی مباحث تشخیصی پژوهش‌گران تحقیقات وسیعی جهت بسط این مباحث در شرایط وجود هم‌خطی در مدل انجام دادند.

[1] Mean Square Error Matrix (MSEM)
[2] Diagnostic Methods
[3] Outliers
[4] Influential observations
[5] Prediction matrix
[6] Cook's Distance
[7] Mean shift

۱-۲ تاریخچه‌ی پژوهش‌های پیشین

آنالیز رگرسیونی در مباحث آماری پیشینه‌ای طولانی دارد. در آغاز مفهوم رگرسیون در نظریه‌ی کمترین توان‌های دوم که توسط برخی محققین طراحی شده بود، ظهور پیدا کرد. اصطلاح رگرسیون به وسیله‌ی گالتون[1] (۱۹۸۹؛۱۸۷۷) برای توصیف یک مسأله‌ی زیست‌شناختی ابداع گردید. برای گالتون رگرسیون مفهومی زیست‌شناختی داشت، اما یول[2] (۱۸۹۷) و پیرسون[3] (۱۹۰۳) کارهای او را برای مفاهیم آماری توسعه دادند. بعدها فیشر[4] (۱۹۲۲) با اشاره به ضعیف بودن فرض‌های آن‌ها این مفهوم را به طور گسترده‌تری مطرح ساخت. امروزه رگرسیون با داشتن نتایج مطلوب یکی از روش‌های پرکاربرد جهت آنالیز داده‌ها محسوب می‌گردد.

با توجه به این موضوع که وجود هم‌خطی در میان متغیرهای مستقل نتایج حاصل از تحلیل رگرسیونی را تحت تأثیر قرار می‌دهد باید جهت اصلاح این مشکل برآییم. مفهوم هم‌خطی، روش‌های تشخیص و برخورد با این مشکل و همچنین پیامدهای ناشی از آن به‌طور وسیعی توسط مونتگمری و پک[5]، (۱۹۸۲) بحث شده است. راه‌حل‌های زیادی برای مقابله با این مسأله ارائه شده است. توجه به کلاس برآوردگرهای اریب و بهره‌مندی از آن‌ها یکی از راه‌حل‌های متداول برای مقابله با این مشکل است.

[1] Galton
[2] Yule
[3] Pearson
[4] Fisher
[5] Montgomery and Peck

۱-۲-۱ برآوردگرهای اریب

استاین[1] (۱۹۵٦) برآورد اریبی جهت رفع مشکل هم‌خطی معرفی کرد. برآورد اریب دیگری تحت عنوان برآورد ریج توسط هورل و کنارد[2] (۱۹۷۰ b,a) مطرح شد. این برآوردگر یکی از روش‌های رایج به منظور تقلیل اثرات هم‌خطی است. به همین جهت جایگاه ویژه‌ای در میان شیوه‌های رفع این مسأله داشته و توجه بسیاری از آماردانان به این موضوع جلب شده است. دمپستر[3] (۱۹۷۷) و وینود[4] (۱۹۷۸) نیز برآورد ریج را مورد بررسی قرار داده و مطالبی را در این زمینه ارائه کردند. روش رگرسیون ریج توسط برخی محققان بسط داده شد و گونه‌های مختلف از این روش ارائه گردید. هورل و کنارد (۱۹۷۰) گسترشی را در روش رگرسیون ریج پیشنهاد کردند به این صورت که برای هر متغیر رگرسیونی پارامتر اریب[5] جداگانه‌ای را اجازه می‌دهد. این روش به عنوان رگرسیون ریج تعمیم یافته[6] شناخته می‌شود. رگرسیون ریج وزنی[7] نیز توسط هولند[8] (۱۹۷۳) مطرح گردید. ترنکلر[9] (۱۹۸٤) برآورد ریج را برای مدلی با خطاهای وابسته و غیر همگن به دست آورد. اسویندل[10] (۱۹۷٦) برآورد ریج اصلاح شده[11] را با استفاده از یک بردار اطلاع پیشین معرفی کرد و پلیسکین[12] (۱۹۸۷) به مقایسه‌ی این برآورد با برآورد ریج پرداخت. برآورد ریج نااریب[13] توسط سینگ[14] (۱۹۸٦) ارائه شد. نسخه‌ی دیگری از برآورد ریج نااریب توسط کراس و

[1] Stein
[2] Horel and Kennard
[3] Dempster
[4] Vinod
[5] Bias parameter
[6] Generalized Ridge regression
[7] Weighted Ridge Regression
[8] Holland
[9] Trenkler
[10] Swindel
[11] Modified Ridge Estimator
[12] Pliskin
[13] Unbiased Ridge Estimator
[14] Singh

همکاران[۱] (۱۹۹۵) مطرح گردید. به دنبال آن فلاح و سلام[۲] (۲۰۱۱) برآورد ریج نااریب اصلاح شده[۳] را معرفی نموده و در مورد ویژگی‌های آن بحث کردند. مسی[٤] (۱۹٦۵) برآورد اریب مؤلفه-های اصلی[۵] را ارائه کرد. لیو[٦] (۱۹۹۳) برآورد جدیدی با ادغام برآورد ریج و استاین و به منظور ترقی این دو برآورد پیشنهاد داد. آکدنیز و کسیرنلر[۷] (۱۹۹۵) این برآورد را با لفظ برآورد لیو معرفی کردند و ساکالی‌اُلگو و همکاران[۸] (۲۰۰۱) نیز ماتریس میانگین توان دوم خطای برآورد ریج و لیو را باهم قیاس کرد. همچنین آکدنیز و ارول[۹] (۲۰۰۳) پس از بررسی برخی برآوردگرهای اریب ذکر شده به مقایسه‌ی آن‌ها با یکدیگر اقدام کرد.

۱-۲-۲ محدودیت‌های خطی

گاهی اطلاعاتی راجع به برخی ضرایب رگرسیونی در دسترس است. برخی آماردانان چون بلسلی و همکاران[۱۰] (۱۹۸۰) و رائو و توتنبرگ[۱۱] (۱۹۹۵) رائو و همکاران (۲۰۰۸) بیان کردند که اضافه کردن این اطلاعات به نمونه موجب بهبود فرایند برآورد ضرایب می‌شود. همچنین آن‌ها فرض کردند که در طی فرایند، اطلاعات پیشین و مشاهدات، وزن یکسانی را دریافت می‌کنند. در عمل موقعیت‌هایی ممکن است اتفاق بیفتد که چنین فرضی امکان‌پذیر نباشد. سیبر و لی[۱۲] (۲۰۰۲) و تیل[۱۳] (۱۹٦۳) آزمون آماری جهت بررسی سازگاری و معنی‌دار بودن این اطلاعات به فرم

[1] Crouse et al
[2] Fallah and salam
[3] Modified Unbiased Ridge Estimator
[4] Massy
[5] Principal Components Regression
[6] Liu
[7] Akdeniz and Kaciranlar
[8] Sakallioglu et al
[9] Akdeniz and Erol
[10] Belsley et al
[11] Rao and Toutenburg
[12] Seber and Lee
[13] Theil

محدودیت‌های دقیق و یا تصادفی ارائه دادند. آن‌ها بیان کردند که در صورت سازگاری، ترکیبی از اطلاعات نمونه با این اطلاعات به فرم محدودیت‌های خطی برای برآورد ضرایب به‌کار برده می‌شود. گاهی معلومات موجود دقیق نبوده و با اندکی خطا همراه می‌باشند، از طرفی نادیده گرفتن این اطلاعات عملکرد ناصحیحی است. بنابراین تیل و گولدبرگر[1] (۱۹۶۱) و تیل (۱۹۶۳) به‌کارگیری این معلومات به فرم محدودیت‌های خطی تصادفی را پیشنهاد کردند.

۱-۲-۳ ترکیب برآوردهای اریب و محدودیت‌های خطی

راه دیگر پاسخ‌گویی به مسأله‌ی هم‌خطی استفاده از اطلاعات کمکی موجود افزون بر نمونه است. این روش توسط بلسلی و همکاران (۱۹۸۰) مطرح گردید. آن‌ها بیان کردند که استفاده از این اطلاعات به دو شیوه‌ی بیزی و برآورد آمیخته امکان‌پذیر است. استفاده از روش بیزی برای بهبود شرایط حاکم توسط زلنر[2] (۱۹۷۱) و لیمیر[3] (۱۹۷۳، ۱۹۷۸) بررسی شده است. مشکل اصلی این روش این است که نیاز به شرح دقیقی از توزیع پیشین دارد. روش برآورد آمیخته مشابه با روش بیزی است. اولین بار دوربین[4] (۱۹۵۳) به‌طور هم‌زمان اطلاعات پیشین و نمونه را برای بهبود برآورد به‌کار گرفت. سپس برآورد آمیخته به وسیله‌ی تیل و گولدبرگر (۱۹۶۱) و تیل (۱۹۶۳) معرفی شد. در این روش اطلاعات کمکی به‌طور مستقیم در قالب محدودیت‌های خطی تصادفی به مجموعه‌ی داده‌ها اضافه می‌شود و این شیوه نیاز به توزیع پیشین ندارد. سارکار[5] (۱۹۹۲) ایده‌ی ترکیب برآورد ریج و برآورد کمترین توان‌های دوم محدود شده[6] را ارائه نموده و در پی آن برآورد

[1] Theil and Goldberger
[2] Zellner
[3] Leamer
[4] Durbin
[5] Sarkar
[6] Restricted least-Squares Estimator

ریج محدود شده[1] را معرفی کرد. وجه تمایز و برتری برآورد ریج اصلاح شده نسبت به برآورد ریج و برآورد ریج محدود شده توسط کسیرنلر و همکاران (۱۹۹۸) و ویجکون[2] (۱۹۹۸) نشان داده شد. کسیرنلر و همکاران (۱۹۹۹) دیدگاه برآورد لیو اصلاح شده را از طریق متحد کردن برآورد لیو و کمترین توان‌های دوم محدود شده ارائه دادند و برآورد لیو محدود شده [3]را معرفی کردند. گروب[4] (۲۰۰۳) بیان کرد که برآورد پیشنهادی سارکار (۱۹۹۲) برای کلیه‌ی محدودیت‌های های خطی در نظر گرفته شده برآوردی رضایت بخش نیست. به همین منظور پس از مطالعه و بررسی‌های لازم برآورد ریج محدود شده‌ی تازه‌ای را که از ادغام روش کمترین توان‌های دوم محدودشده با برآورد پیشنهادی اسویندل (۱۹۷٦) حاصل شده بود ارائه داد. هابرت[5] و جیکون (۲۰۰٦) برآورد لیو تحت محدودیت تصادفی[6] را با ایجاد اصلاحاتی در برآورد لیو محدود شده پیشنهاد دادند این برآورد تعمیمی از برآورد لیو و آمیخته است. ریوان اوزکال[7] (۲۰۰۹) نظریه‌ی بهره‌مندی هم‌زمان از رگرسیون ریج و محدودیت‌های خطی تصادفی را ارائه کرد و همچنین برآورد ریج آمیخته[8] را برای موقعیتی که در آن بردار خطای مدل همبسته و غیر همگن باشد به دست آورد. نسخه‌ی دیگری از برآورد ریج آمیخته در سال‌های اخیر توسط لی و یانگ[9] (۲۰۱۰) پیشنهاد و مورد مطالعه قرار گرفته است.

[1] Restricted Ridge Estimator
[2] Wijekoon
[3] Restricted Liu Estimator
[4] Grob
[5] Hubert
[6] Stochastic Restricted Liu Estimator
[7] Revan Ozkale
[8] Stochastic Restricted Ridge Estimator
[9] Li and Yang

۱-۲-٤ آنالیز تشخیصی

در تحلیل رگرسیون مفروضات عمده‌ای در نظر گرفته می‌شود. برای آزمون مناسب بودن مدل باید به اعتبار این مفروضات توجه داشت. انحرافات شدید از مفروضات می‌تواند به مدل ناپایدار منجر شود. برای تشخیص و برخورد با انحراف از این مفروضات، روش‌های مختلفی پیشنهاد شده است که مجموعه‌ی این روش‌ها را به عنوان مباحث تشخیصی بیان می‌کنیم (مونتگمری و پک، ۲۰۰۱). مطالعه‌ی مباحث تشخیصی به ویژه یافتن مشاهدات پرت و تأثیرگذار در مدل‌های رگرسیونی از اهمیت وافری برخوردار است. نویسندگان و محققان زیادی در این زمینه به تحقیق و بررسی پرداختند. به‌طور مشخص تعیین مشاهدات مؤثر و پرت توجه بسیاری از آماردانان از جمله کوک[۱] (۱۹۷۷)، بلسلی و همکاران (۱۹۸۰)، کوک و ویسبرگ[۲] (۱۹۸۲)، چترجی و هادی[۳] (۱۹۸٦ و ۱۹۸۸) و آتکینسون[٤] (۱۹۸۸) را جلب کرده است. بلسلی و همکاران (۱۹۸۰) بیان کردند که وجود هم‌خطی در مدل می‌تواند حالت هم‌پوشانی برای این مشاهدات ایجاد کند. همچنین بیان شد که در صورت پدیدار شدن هم‌خطی ممکن است مشاهدات تأثیرگذار در برآورد ریج متمایز با برآورد کمترین توان‌های دوم باشد (والکر و بیرچ[٥] ۱۹۸۸). روش‌های تشخیص این مشاهدات برای رگرسیون ریج توسط والکر و بیرچ (۱۹۸۸) ارائه شد. در مقاله‌ی ارائه شده توسط استیک[٦] (۱۹۸٦) ذکر شد که اثر مشاهدات بانفوذ[۷] بر برآورد ریج کمتر است از اثری که همان مشاهدات بر برآورد کمترین توان‌های دوم دارند. رویکرد تأثیر مکانی[۸] تعمیمی از روش‌های تشخیصی است که توسط

[1] Cook
[2] Weisberg
[3] Chatterjee and Hadi
[4] Atkinson
[5] Walker and Birch
[6] Steece
[7] Leverage Observations
[8] Local Influence

توسط کوک (۱۹۸٦) ارائه شد. شی[1] (۱۹۹۷) و شی و وانگ[2] (۱۹۹۹) به ترتیب به بررسی آنالیز تأثیر مکانی در روش مؤلفه‌های اصلی و روش رگرسیون ریج پرداختند. روش‌های تشخیصی موردی[3] برای رگرسیون ریج اصلاح شده توسط جاهوفر و چن[4] (۲۰۰۹) بسط داده شد. جاهوفر و جاهوفر و چن (۲۰۱۱) به مطالعه‌ی آنالیز تأثیر مکانی در رگرسیون ریج اصلاح شده پرداختند.

۱-۳ ساختار پایان‌نامه

تشخیص مشاهدات تأثیرگذار به دلیل اثرات نامناسب و غیرمعقولی که بر نتایج حاصل از تحلیل رگرسیونی دارند در مباحث آماری از اهمیت ویژه‌ای برخوردار است. بنابراین در این رساله بر مطالعه‌ی مباحث تشخیصی چون یافتن مشاهدات تأثیرگذار و داده‌های پرت در شرایط حضور هم‌خطی و استفاده از رگرسیون ریج تمرکز می‌شود و در ادامه نتایج به دست آمده را برای رگرسیون ریج تحت محدودیت‌های خطی تصادفی تعمیم می‌دهیم. با این هدف در فصل اول مقدمه و تاریخچه‌ای در این زمینه ارائه گردیده است. فصل دوم این رساله، به بیان پدیده‌ی هم-خطی و مشکلات حاصل از بروز آن می‌پردازد. همچنین مفصلاً روش‌های شناسایی هم‌خطی شرح داده می‌شود. در ادامه فصل، به بیان روش رگرسیون ریج و برآورد حاصل از این روش خواهیم پرداخت، همچنین ویژگی‌های برآورد ریج را برمی‌شماریم. در فصل سوم استفاده از اطلاعات اضافی به فرم محدودیت‌های خطی با جزییات بیشتر شرح داده می‌شود. به دنبال آن برآورد محدود شده و برآورد آمیخته معرفی می‌گردد. در پی آن نیز برآورد ریج تحت محدودیت دقیق و برآورد ریج تحت محدودیت تصادفی و خواص آن شرح داده می‌شود. فصل چهارم مفهوم مباحث تشخیصی را تشریح کرده و برخی معیارهای تشخیصی را جهت شناسایی نقاط تأثیرگذار

[1] Shi
[2] Wang
[3] Case deletion
[4] Jahufer and Chen

معرفی می‌کنیم. در ادامه بررسی‌های انجام شده در جهت تخصیص این معیارها برای رگرسیون ریج را بیان کرده و معیارهای تشخیصی تعمیم یافته را شرح خواهیم داد. فصل پنجم، روش‌های تشخیص مشاهدات تأثیرگذار و پرت تحت شرایط استفاده هم‌زمان از رگرسیون ریج و اعمال محدودیت‌های خطی تصادفی ارائه شده و همچنین نتایج تئوری بیان شده در این فصل و فصل-های گذشته را با ارائه‌ی یک مثال عددی تشریح می‌کنیم و در پایان نتایج و پیشنهادات را مطرح خواهیم کرد.

فصل دوم

هم‌خطی و برآورد ریج

۲-۱ مقدمه

مدل رگرسیونی خطی بیان کننده‌ی یک رابطه‌ی خطی بین متغیرهای مستقل و متغیر پاسخ است. چه بسا رابطه‌ی قوی تجربی بین دو یا چند متغیر وجود داشته باشد؛ اما دلیلی بر وجود رابطه‌ی علت و معلولی بین متغیرها نیست. تحلیل رگرسیونی می‌تواند در تصدیق و تأیید این رابطه کمک کننده باشد. در تحلیل و به‌کارگیری مدل رگرسیونی برآورد ضرایب یکی از اهداف و استنباط‌های رایج است؛ که محققان به منظور نائل آمدن به این هدف بسیار کوشیده‌اند. یکی از چالش‌های پیش‌رو در این زمینه مشکل هم‌خطی است. بدین منظور در اینجا به طرح این معضل پرداخته و بیان می‌کنیم که وجود هم‌خطی در میان متغیرهای مستقل چه مشکلاتی را به همراه خواهد داشت. همچنین شاخص‌هایی را که برای تشخیص هم‌خطی پیشنهاد شده ارائه می‌دهیم. از طرفی برای رفع مشکل هم‌خطی راه‌حل‌هایی مطرح شده است که در این فصل به بیان یکی از این روش‌ها یعنی به‌کارگیری کلاس برآوردگرهای اریب و به طور ویژه برآورد ریج می‌پردازیم. در ادامه ویژگی‌های برآورد حاصل را ذکر کرده، در نهایت به مقایسه‌ی برآورد ریج و برآورد کمترین توان-های دوم می‌پردازیم.

۲-۲ هم‌خطی

در آغاز پیش از طرح موضوع بنا به ضرورت و پیش‌برد رساله، یک مدل رگرسیون خطی به عنوان پیش فرض قرار می‌دهیم؛ بنابراین مدل رگرسیونی زیر را در نظر می‌گیریم.

$$Y = X\beta + \varepsilon \qquad , \ \varepsilon \sim N(0, \sigma^2 I) \tag{۱.۲}$$

که در آن Y یک بردار $n \times 1$ از مشاهدات، X یک ماتریس $n \times p$ از متغیرهای مستقل با رتبه‌ی p، β یک بردار $p \times 1$ از ضرایب رگرسیونی نامعلوم و ε یک بردار $n \times 1$ از خطاهاست. برآورد ضرایب را با استفاده از روش متداول کمترین توان‌های دوم بر مبنای کمینه کردن مجموع توان دوم خطا به‌صورت زیر محاسبه می‌کنیم.

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

در مدل‌های رگرسیونی با ماتریس متغیرهای مستقل پر رتبه، برآورد ضرایب رگرسیونی به سادگی به‌دست می‌آید؛ اما در مواردی علیرغم مستقل بودن این متغیرها به دلیل وابستگی زیاد آن‌ها برآوردهای حاصل قابل اعتماد نخواهند بود و در مواردی می‌توانند به استنتاج نادرستی منجر شوند.

موقعیتی را که در آن بین متغیرهای مستقل وابستگی خطی وجود داشته باشد تحت عنوان مسأله‌ی هم‌خطی مطرح می‌کنند. اغلب بین هم‌خطی کامل و ناقص تمایز وجود دارد. وقتی که برخی از ستون‌های ماتریس متغیرهای مستقل ارتباط خطی دقیق و کامل داشته باشند آنگاه هم‌خطی کامل خواهیم داشت. به بیان دیگر می‌توان هم‌خطی کامل را به این صورت تعریف کرد که اگر برای زیر مجموعه‌ای از ستون‌های X رابطه‌ی زیر برقرار باشد که در آن t_j مقادیر ثابت غیر صفرند و X_j نشان دهنده‌ی j امین ستون X است، آنگاه در مدل هم‌خطی کامل وجود دارد.