دانشکده علوم ریاضی و کامپیوتر

گروه آمار

پایان نامه کارشناسی ارشد آمار

گرایش محض

عنوان:

تأثیر مشاهدات بر آماره‌های آزمون با تمرکز روی مدل‌های رگرسیون خطی

نگارش:

فاطمه قنداق‌ساز

اساتید راهنما:

دکتر عبدالرحمن راسخ

دکتر قاسم تارمست

استاد مشاور:

دکتر صادق رضایی

تیر ۹۰

بسمه تعالی

«چکیده پایان‌نامه»

نام خانوادگی دانشجو: قنداق‌ساز **نام**: فاطمه

عنوان پایان نامه: تأثیر مشاهدات بر آماره های آزمون با تمرکز روی مدل های رگرسیون خطی

اساتید راهنما: دکتر عبدالرحمن راسخ و دکتر قاسم تارمست

درجه تحصیلی: کارشناسی ارشد **رشته:** آمار **گرایش**: محض

محل تحصیل (دانشگاه): شهید چمران اهواز

دانشکده: علوم ریاضی و علوم کامپیوتر

تاریخ فارغ التحصیلی: ۱۳۹۰/۰٤/۰٦

تعداد صفحات: ۱۳۲

کلیدواژه‌ها: آماره‌ی نسبت درستنمایی، آنالیز تشخیص حذف موردی، باقی‌مانده‌های استیودنت شده، تابع تأثیر، ماتریس تصویر.

چکیده: از جمله مسائلی که در روش‌های رگرسیونی از اهمیت ویژه‌ای برخوردار است، بررسی میزان تأثیری است که هر مشاهده بر هر یک از جنبه‌های مدل‌بندی دارد. اما از آن‌جا که تأثیر هر مشاهده بر هر یک از جنبه‌های رگرسیونی یکسان نیست، شناسایی مشاهدات مؤثر در هر زمینه مستلزم مطالعه‌ی تأثیر در همان زمینه خاص می‌باشد. هدف اساسی این پایان‌نامه بررسی آنالیز تشخیص حذف موردی برای ارزیابی تأثیر هر مشاهده بر آماره‌های آزمون است. در این راستا ابتدا مفاهیم مورد استفاده در پایان‌نامه را مرور می‌کنیم، سپس به مطالعه‌ی تابع تأثیر آزمون‌ها در مدل-های رگرسیونی خطی پرداخته و برای این مدل‌ها، دو فرم تابع تأثیر نسبتی و تفاضلی آماره‌ی آزمون را بررسی می‌کنیم. سپس این توابع تأثیر را برای مدل‌های رگرسیونی وزنی به‌دست آورده و با استفاده از آن‌ها مشاهدات را براساس میزان تأثیری که در آزمون‌ها دارند، رتبه‌بندی و مؤثر بودن هر مشاهده در آماره‌ی آزمون را مورد سنجش و ارزیابی قرار می‌دهیم. سرانجام برای تأیید مباحث نظری ارائه شده، مثال‌هایی از داده‌ای واقعی را مطرح می‌کنیم.

فهرست مطالب

عنوان صفحه

فصل اول: مقدمه و کلیات

فصل دوم: معرفی مدل‌های رگرسیونی

فصل سوم: ماتریس تصویر و انواع باقی‌مانده‌ها

فصل چهارم: تابع تأثیر و آزمون محدودیت‌های خطی

فصل پنجم: تابع تأثیر آماره آزمون‌های نسبت درست‌نمایی در مدل‌های رگرسیونی

فصل ششم: تابع تأثیر آزمون‌های مدل‌های رگرسیونی وزنی

نتیجه‌گیری و پیشنهادات

فصل اول:

مقدمه و کلیات

۱-۱ مقدمه

یکی از مسائلی که از دیرباز ذهن بشر را به خود مشغول کرده است، شناسایی پدیده‌های طبیعی و فیزیکی و تلاش برای معرفی آن‌ها می‌باشد. زمانی که عوامل مؤثر در تشکیل یک پدیده قابل اندازه‌گیری باشند، یکی از ساده‌ترین راه‌ها برای معرفی رفتار آن پدیده، استفاده از یک مدل ریاضی است. گاهی برای سادگی بیشتر ممکن است آگاهانه یا ناآگاهانه، با چشم‌پوشی از بررسی برخی فاکتورهای مهم، زمینه ایجاد نوعی خطا در مدل فراهم گردد و یا حتی زمانی که با اطمینان کامل تمامی فاکتورهای مهم در مدل گنجانده شوند، باز هم به دلیل خطاهای موجود در اندازه - گیری‌ها وعدم دقت کافی، مدل‌بندی با نوعی خطای غیر قابل اندازه‌گیری همراه خواهد بود. این مسئله در عمل ما را به استفاده از مدل‌های آماری رهنمون می‌سازد که این عدم اطمینان با اضافه

کردن عبارت خطا در مدل گنجانده می‌شود. ضمن این‌که با استفاده از این مدل‌ها امکان بررسی و تجزیه و تحلیل‌های آماری، میسر می‌شود (رایان،[۱] ۱۹۹۷).

۱-۲ مدل رگرسیونی خطی

ساده‌ترین مدلی که به کمک آن می‌توان رابطه‌ی یک متغیر مشاهده شده را با p متغیر دیگر توصیف کرد، مدل رگرسیونی خطی[۲] است. این مدل که بر حسب پارامترها خطی است، به صورت زیر تعریف می‌شود:

$$Y = X\beta + \varepsilon \qquad (۱-۱)$$

که در آن Y، بردار$n \times ۱$ از متغیر پاسخ، X ماتریس $n \times p$ از متغیرهای توضیحی، β بردار p بعدی از پارامترهای براوردپذیر و ε بردار n بعدی عبارت خطا است. در عمل مدل‌های خطی کاربرد بیشتری نسبت به سایر مدل‌ها دارند؛ زیرا علاوه بر سادگی محاسبات، به راحتی نیز قابل تفسیرند. همچنین بسیاری از مدل‌های غیرخطی را می‌توان با کمک تبدیل‌هـــای ساده ریاضی به مدل‌های خطی تبدیل کرد.

۱-۳ روش‌های آنالیز تشخیص

هنگام استفاده از مدل‌های به‌فرم (۱-۱)، مدل‌بندی بر مبنای یک سری فرضیات صورت می‌گیرد و اعتبار استنباط‌های ناشی از این مدل‌ها، مستلزم صحت این فرضیات است. در دهه‌های اخیر تأکید فراوانی بر ارزیابی فرض‌های آماری صورت گرفته و مطالعات گسترده‌ای در این زمینه انجام شده است. همچنین به‌منظور بررسی تأثیر مشاهدات بر براوردهای پارامترها، مقادیر پیش‌بینی و نیز جنبه‌های دیگر مدل روش‌هایی تحت عنوان آنالیز تشخیص پیشنهاد شده است. اهمیت استفاده از آنالیز تشخیص در آن است که علاوه بر شناسایی جنبه‌هایی از مدل که با فرضیات مدل‌بندی هماهنگ نیستند، برای اصلاح آنالیز مدل نیز ایده‌های مطلوبی ارائه می‌دهد و به کمک آن

[۱] Ryan
[۲] Linear Regression Model

می‌توان به این پرسش پاسخ داد که؛ چگونه می‌توان به بررسی ثبات یا مطالعه انحراف در آنالیز نتایج پس از تعیین یک فرمول‌بندی پیچیده پرداخت؟ (کوک و ویزبرگ[۱]، ۱۹۸۲)

کوک و ویزبرگ (۱۹۸۲) مطالعه وابستگی نتایج و استنباط‌ها به جنبه‌های مختلف فرمول‌بندی یک مسئله را مطالعه‌ی تأثیر[۲] نامیدند. ایده اصلی آنالیز تأثیر بسیار ساده و به این صورت است که ابتدا یک اغتشاش[۳](پرشیدگی) جزئی در فرمول‌بندی‌های پیچیده اعمال می‌شود و سپس بررسی می‌شود که این اغتشاش چگونه نتایج آنالیز را تغییر می‌دهد؟ یکی از انواع اغتشاشات که بیشترین مقبولیت را نسبت به سایر روش‌ها دارا می‌باشد، روش حذف موردی[۴] است که در آن پس از حذف هر مشاهده، تأثیر ایجاد شده توسط آن مشاهده بررسی می‌شود (کوک و ویزبرگ، ۱۹۸۲).

یکی از فرضیاتی که مدل‌بندی بر مبنای آن صورت می‌گیرد، عدم وجود داده‌ی مؤثر است. اما این فرض به‌سختی برقرار است و معمولاً نقش همه‌ی مشاهدات در تعیین براوردها، آماره‌های آزمون و دیگرآماره‌ها یکسان نیست، به‌طوری که در برخی مسائل، ویژگی‌های رگرسیونی تنها تابع وجود مشاهداتی خاص بوده وعملاً از بیشتر مشاهدات چشم‌پوشی می‌شود. به همین خاطر شناسایی مشاهدات مؤثر و بررسی اثر آن‌ها بر روی جنبه‌های مختلف یک آنالیز، برای تحلیلگر از اهمیت ویژه‌ای برخوردار است. ضمن مطالعه تأثیر، علاوه بر شناسایی کامل این گونه مشاهدات، اطلاعاتی در رابطه با میزان اطمینان به نتایج حاصل و وابستگی آن‌ها به مدل مفروض در اختیار آماردان قرار می‌گیرد. به‌عنوان نمونه، اگر حذف یک مشاهده‌ی مؤثر از مجموعه‌ی داده‌ها علامت یک پارامتر براورد شده را تغییر دهد، به کار بردن استنباط‌های مربوط به آن پارامتر با تردید همراه خواهد بود (کوک و ویزبرگ، ۱۹۸۲).

علاوه بر این، با مطالعه‌ی تأثیر، بخش‌هایی که در آن‌ها داده‌ی گمشده وجود دارد و در آن قسمت از فضای مشاهدات پوشش کافی برای براورد و پیش‌بینی وجود ندارد، مشخص می‌شود. اگر مشاهدات مؤثر و یا مشاهدات دورافتاده ناشی از خطای اندازه‌گیری یا شرایط نامناسب

---

[۱] Cook & Weisberg
[۲] Study of influence
[۳] Perturbation
[۴] Case Deletion

آزمایشگاهی باشند، باید حذف شوند در غیر این صورت جمع‌آوری داده‌های بیشتر اقدامی مناسب خواهد بود (کوک و ویزبرگ، ۱۹۸۲).

۱-٤ مطالعه تأثیر

مطالعه‌ی تأثیر براساس هر یک از پنج معیار زیر می‌تواند صورت گیرد:

۱- باقی‌مانده‌ها

۲- ماتریس تصویر

۳- اندازه‌ی بیضی‌های اطمینان

٤- تابع تأثیر

۵- تأثیر موضعی

که از میان این معیارها، بررسی باقی‌مانده‌های مدل برازش شده، از مهم‌ترین ابزارهای آنالیز تشخیص است که به کمک آن می‌توان درستی فرضیاتی چون وجود داده‌ی دورافتاده و یا همگنی واریانس[۱] را تحقیق کرد. یک داده‌ی دورافتاده در واقع مشاهده‌ای است که به طرز آشکاری از مجموعه‌ی اکثریت داده‌ها به دور است و ممکن است نتایج برازش مدل را به مقدار زیادی تحت تأثیر قرار دهد به گونه‌ای که با حذف آن مشاهده از مجموعه داده‌ها نتایج کاملاً متفاوتی به‌دست می‌آید. یک ملاک برای شناسایی مشاهدات دور افتاده بررسی باقی‌مانده‌های مدل است. ضمن این‌که مواردی نیز وجود دارند که داده دورافتاده نبوده، اما نتایج کاملاً به وجود آن مشاهدات وابسته است که آن‌ها را مشاهدات مؤثر[۲] می‌نامند. به این ترتیب نه تنها مشاهدات دور افتاده بلکه هر داده‌ای ممکن است مؤثر تلقی شده و نتایج آنالیزی را تا حد زیادی، تحت تأثیر خود قرار دهد (چترجی و هادی[۳]، ۱۹۸٦).

[۱] Homoskedasicity of variance
[۲] Influential observations
[۳] Chatterjee&Hadi

با توجه به این‌که هدف اصلی این مطالعه، شناسایی این گونه مشاهدات غیر معمول است، آنالیز باقی‌مانده‌ها به تنهایی برای شناسایی این موارد کافی نیست. بدین جهت در این پایان‌نامه، علاوه بر مطالعه‌ی باقی‌مانده‌ها به مطالعه‌ی تابع تأثیر نیز می‌پردازیم.

تابع تأثیر از جمله معیارهای مطالعه‌ی تأثیر است که تحلیلگر به کمک آن می‌تواند مشاهدات را براساس میزان تأثیرشان رتبه‌بندی کند. ایده‌ی اصلی تابع تأثیر اولین بار توسط همپل[۱](۱۹۷٤) مطرح شد و طرح کلی آن به این صورت است که رفتار جزئی توابع آماری مختلف و نیز سهم تک تک مشاهدات بر روی آن‌ها مورد بررسی قرار گرفته و با توجه به آن، به هر مشاهده رتبه‌ای اختصاص می‌یابد، سپس مشاهده‌ای که بیشترین تأثیر را داشته باشد، به عنوان مؤثرترین مشاهده معرفی می‌شود (همپل، ۱۹۷٤).

۱- ۵ تاریخچه‌ی اولیه

هر مشاهده از مدل مورد نظر می‌تواند تأثیر زیادی در براورد پارامترها و به دنبال آن در استنباط‌هایی که براساس این براوردها صورت می‌گیرد، داشته باشد. در زمینه‌ی شناسایی مشاهدات مؤثر در براوردها، مقالات زیادی ارائه شده است که سرآغاز همه‌ی آن‌ها مقاله‌ای بود که توسط کوک (۱۹۷۷) ارائه شد و در آن آنالیز تشخیص حذف موردی برای مدل‌های خطی براساس تفاوت بین براوردهای پارامتری حاصل از مجموعه‌ی داده‌های کامل و مجموعه‌ی داده-هایی که یک مشاهده از آن‌ها حذف شده، تحت عنوان فاصله‌ی کوک[۲] معرفی شد (ریتز و اسکوگارد،[۳] ۲۰۰۷).

به دنبال آن روش‌های متعددی برای مدل‌های مختلف به دست آمد. به عنوان مثال کوک (۱۹۷۹) به بررسی تأثیر هر مشاهده در مدل‌های رگرسیونی خطی پرداخت. پرگیبان[٤] (۱۹۸۱) ابزارهای آنالیز تشخیص را برای رگرسیون لجستیک مطرح کرد. ویلیامز (۱۹۸۷) مباحث آنالیز

[۱] Hample
[۲] Cook's Distance
[۳] Ritz & Skogard
[٤] Pregibon

تشخیص حذفی را برای مدل‌های خطی تعمیم یافته[۱] گسترش داد. کریستن‌سن[۲] و همکاران (۱۹۹۲) روشی برای مدل‌های آمیخته خطی[۳] ارائه دادند. راسن[٤] (۱۹۹۵) به بررسی مشاهدات مؤثر در مدل‌های رگرسیونی چندمتغیره پرداخت. شی و چن[٥] (۲۰۰۸) آنالیز تشخیص حذفی در مدل‌های چندسطحی را بررسی کردند.

آن‌چه مسلم است، این است که تأثیر یک مشاهده روی همه‌ی کمیت‌های رگرسیونی یکسان نیست و شناسایی نقاط مؤثر در هر زمینه، مستلزم مطالعه‌ی تأثیر، در آن زمینه است و این‌گونه نیست که اگر یک مشاهده مثلاً در براورد پارامترها مؤثر باشد، بر آماره‌ی آزمون نیز مؤثر باشد. آنالیز تشخیص برای بررسی تأثیر یک مشاهده روی آزمون فرض‌ها چندان مورد توجه قرار نگرفته و مطالعات اندکی در این زمینه صورت گرفته است. به‌عنوان نمونه تغییرات آماره‌ی F جزئی ابتدا در کوک (۱۹۷۹) و سپس به‌صورت کلی‌تر، در کوک و ویزبرگ (۱۹۸۲) مورد بررسی قرار گرفت. لاس‌بدر و مول‌گاکر[٦](۱۹۸۵) به شناسایی نقاط مؤثر در آزمون‌های امتیاز با استفاده از آنالیز تشخیص حذفی پرداختند. ویلیامز[۷] (۱۹۸۷) تقریب یک مرحله‌ای را برای تغییر در آماره‌ی نسبت لگاریتم درست‌نمایی محاسبه کرد. جالیف و لاکادا[۸] (۱۹۹۳) تأثیر تک مشاهدات را بر روی آزمون‌های تک مشاهده‌ای ساده برای میانگین و واریانس مورد مطالعه قرار دادند. گرمیت و ریدن-آور[۹] (۱۹۹٦) اثر تک مشاهدات را بر روی آزمون‌های معمولی میانگین بررسی کردند. لی و ژائو[۱۰] ژائو[۱۰] (۱۹۹٦ و ۱۹۹۷) تأثیر مشاهدات را بر آماره‌ی نیکویی برازش پیرسن و نیکویی تابع پیوند در مدل‌های خطی تعمیم یافته مطالعه کردند. کیم[۱۱] (۱۹۹۸) روش تأثیر موضعی را برای مطالعه‌ی تأثیر مشاهدات روی آزمون فرض‌های خطی به‌کار برد و کیم (۲۰۰۲) به مطالعه‌ی تأثیر مشاهدات

[۱] Generalized linear model
[۲] Christensen
[۳] Linear mixed models
[٤] Rosen
[٥] Shi & Chen
[٦] Lusbader & Moolgavkar
[۷] Williams
[۸] Jolliffe & Lukuda
[۹] Grimmett & Ridenhour
[۱۰] Lee & Zhao
[۱۱] Kim

بر روی آزمون فرض‌های خطی با به‌کار بردن آنالیز حذفی موردی یا گروهی پرداخت. ریتز و اسکوگارد (۲۰۰۷) با به‌کار بردن آنالیز تشخیص حذف موردی که در داویسون و اسنل[۱] (۱۹۹۱) تعریف شد، آنالیز حذفی را برای ارزیابی تأثیر یک مشاهده بر روی آزمون‌های نسبت درست‌نمایی بررسی کردند.

در این پایان‌نامه بر روی مدل‌های رگرسیونی وزنی تمرکز می‌کنیم و آنالیز تشخیص حذفی را برای آماره آزمون این مدل‌ها به‌دست می‌آوردیم و با به‌کار بردن آن‌ها مشاهدات را براساس میزان تأثیری که در آماره آزمون دارند، رتبه‌بندی کنیم. اما براساس این معیار قادر به تشخیص مشاهدات مؤثر در آزمون نمی‌باشیم زیرا هیچ مدرکی دال بر این‌که مشاهدات با تأثیر زیاد قطعاً دارای تأثیر غیرمعمول بر آزمون می‌باشد، وجود ندارد! به همین دلیل به این معیار کفایت نکرده و برای حل این مشکل سعی می‌شود تا با تعریف معیارهای تأثیر برای آماره لگاریتم نسبت درست‌نمایی و محاسبه توزیع مجانبی مربوط به آن در مورد مؤثر بودن هر مشاهده در آزمون تصمیم‌گیری شود.

پیکره کلی این پایان‌نامه به این صورت است که در فصل بعد به معرفی مدل‌های رگرسیونی و مدل‌های رگرسیونی وزنی می‌پردازیم و علاوه بر براورد پارامترها، آزمون‌های فرضیه را نیز برای این مدل‌ها مورد مطالعه قرار خواهیم داد. در فصل سوم به تشریح مفاهیم و اصطلاحاتی که در فصل‌های بعد به طور مکرر از آن‌ها استفاده خواهیم کرد، می‌پردازیم. در همین راستا نخست ماتریس تصویر و چگونگی تفسیر آن را بررسی خواهیم کرد. سپس به معرفی انواع باقی‌مانده‌ها برای مدل‌های رگرسیونی خطی و مدل‌های رگرسیونی وزنی می‌پردازیم. در فصل چهارم ضمن تعریف کلی تابع تأثیر، مرور کوتاهی بر بررسی تأثیر هر مشاهده در براورد پارامترها خواهیم داشت و سپس به مطالعه و بررسی تابع تأثیر نسبتی[۲] آماره نسبت درست‌نمایی برای مدل‌های رگرسیونی خواهیم پرداخت. فصل پنجم را با تعریف تابع تأثیر تفاضلی[۳] برای آماره لگاریتم درست‌نمایی برای مدل‌های رگرسیونی خطی آغاز می‌کنیم، پس از آن توجه خود را بر روی تابع

[۱] Davison & Snell
[۲] Ratio influence function
[۳] Difference influence function

توزیع مجانبی این آماره معطوف کرده و از آن برای به‌دست آوردن معیاری برای تصمیم‌گیری در مورد مؤثر بودن هر مشاهده در آزمون، استفاده می‌کنیم. در فصل ششم تابع تأثیر آماره‌های آزمون را در مدل‌های رگرسیونی خطی وزنی به‌دست می‌آوریم. به این ترتیب که ابتدا تابع تأثیر نسبتی را برای این مدل‌ها به‌دست آورده، پس از آن ضمن تعریف تابع تأثیر تفاضلی برای آماره‌ی آزمون این مدل‌ها سعی می‌کنیم توزیع مجانبی آن را نیز به منظور شناسایی مشاهدات مؤثر در آزمون محاسبه کنیم.

فصل دوم:

معرفی مدل‌های رگرسیونی

۲-۱ مقدمه

در این فصل از پایان‌نامه به معرفی مدل‌های رگرسیونی خطی کلی می‌پردازیم و برخی نتایج اساسی در زمینه براوردگرهای کمترین توان‌های دوم تعمیم یافته *(GLSEs)*[۱] را بیان خواهیم کرد. سپس برخی حالات خاص آن شامل مدل‌های رگرسیونی خطی و مدل‌های رگرسیونی وزنی را معرفی کرده و برخی خواص استنباطی در زمینه‌ی براورد پارامترها و آزمون‌های مربوط به این مدل‌ها را بیان خواهیم نمود.

در بخش بعد، فرم کلی مدل‌های رگرسیونی معرفی می‌شوند. پس از آن در بخش سوم به معرفی مدل‌های رگرسیونی خطی با واریانس همگن می‌پردازیم و برای این مدل‌ها، براورد

[۱] Generalized least square estimators

پارامترها، براورد پارامترها تحت محدودیت‌های خطی و آزمون فرض‌های خطی را بیان خواهیم نمود. در بخش آخر بر روی مدل‌های رگرسیونی وزنی تمرکز می‌کنیم. در این مدل‌ها مشاهدات، دارای ساختارهای ناهمگن برای واریانس هستند. هدف این بخش چگونگی مدل‌بندی این داده‌ها و یافتن براوردهای پارامترهای این مدل‌ها در حالت کلی و در حالت محدود شده و انجام آزمون فرض‌های خطی است.

۲-۲ فرم کلی مدل‌های رگرسیونی خطی

مدل رگرسیونی خطی کلی به فرم

$$Y = X\beta + \varepsilon \qquad (۱-۲)$$

که در آن Y یک بردار $n \times ۱$، X یک ماتریس $n \times p$ معلوم پررتبه ستونی و ε بردار تصادفی خطای غیر قابل مشاهده با

$$E(\varepsilon) = ۰ \, , \, cov(\varepsilon) = E(\varepsilon\varepsilon^T) = \Omega \epsilon \mathcal{S}(n)$$

می‌باشد که $\mathcal{S}(n)$ نشان‌دهنده‌ی مجموعه‌ی تمام ماتریس‌های همیشه مثبت $n \times n$ است. اگر $\mathcal{P}_n(۰, \Omega)$ کلاسی گسترده از توزیع‌هایی با میانگین صفر و ماتریس کوریانس Ω تعریف شود، این کلاس به طور خاص توزیع نرمال $N_n(۰, \Omega)$ را در بر می‌گیرد. در این پایان‌نامه، مطالعه‌ی خود را بر روی این توزیع متمرکز می‌کنیم. اما ماتریس کواریانس[۱] Ω معمولاً مجهول است و به عنوان تابعی از یک پارامتر مجهول اما براوردپذیر θ به‌صورت زیر در نظر گرفته می‌شود:

$$\Omega = \Omega(\theta).$$

ساختار کلی ماتریس کواریانس برای عبارت خطای ε، به شکل $\Omega = \sigma^۲ \sum$ است که به وسیله تابعی از یک پارامتر مجهول و براوردپذیر θ مشخص می شود؛ یعنی

$$\sum = \sum(\theta).$$

[۱] Covariance matrix

حالت‌های خاص این مدل، با توجه به ساختارهای خاص ماتریس X و ماتریس کواریانس $\Omega = \sigma^2 \sum(\theta)$ مدل‌های مختلف را نتیجه می‌دهد که هر کدام در شرایط خاص کاربرد دارند. این مدل‌ها به مدل‌های تک متغیره محدود نمی‌شوند، بلکه مدل‌های چند متغیره مانند مدل‌های آنالیز واریانس چند متغیره[۱] *(MANOVA)*، مدل‌های رگرسیونی به‌ظاهر نامرتبط[۲] و مدل منحنی رشد[۳] را نیز در بر می گیرند (کاریا و کوراتا،[۴] ۲۰۰۴).

در این مدل‌ها، برای براورد پارامترها می‌توان از قضیه زیر که به قضیه گوس–مارکوف[۵] معروف است، استفاده کرد:

قضیه (۲–۱). در مدل‌هایی که به فرم (۲–۱) هستند، اگر ماتریس $\sum$ معلوم باشد، براورد پارامترها طبق رابطه‌ی زیر به‌دست می‌آیند:

$$b(\sum) = (\mathrm{X}^{\mathrm{T}}\sum^{-1}X)^{-1}X^{T}\sum^{-1}Y. \qquad (۲-۲)$$

براوردگرهایی که با استفاده از این قضیه به‌دست می‌آیند، بهترین براوردگرهای نااریب خطی[۶] $(BLUE)$ می‌باشند و به براوردهای گوس–مارکوف (GME) معروف‌اند، اما در بیشتر موارد ماتریس $\sum$ مجهول است و بنابراین نمی‌توان از این براوردگرها استفاده کرد. در چنین مواردی از براوردگر کمترین توان‌های دوم تعمیم یافته[۷] $(GLSE)$ استفاده می‌شود که همان GME بوده، با این تفاوت که براوردگر $\widehat{\sum}$ به جای $\sum$ جایگذاری شده است

$$b(\widehat{\sum}) = (\mathrm{X}^{\mathrm{T}}\widehat{\sum}^{-1}X)^{-1}X^{T}\widehat{\sum}^{-1}Y. \qquad (۲-۳)$$

از این براوردگر به طور گسترده برای داده‌های واقعی استفاده می‌شود (کاریا و کوراتا، ۲۰۰۴).

همان‌گونه که بیان شد، مدل (۲–۱) بیشتر مدل‌ها از جمله مدل‌های رگرسیون خطی ساده و مدل‌های رگرسیونی وزنی را نیز در بر می گیرد. در ادامه به معرفی این مدل‌ها می‌پردازیم.

---

۱ Multivariate analysis of variance
۲ Seemingly unrelated regression
۳ Growth curve model
۴ Kariya & Kurata
۵ Gauss-Markov
۶ Best linear unbias estimator
۷ General least square estimator

۲-۳ مدل‌های رگرسیون خطی

در مدل‌های رگرسیون خطی، هدف مورد نظر یافتن رابطه‌ی خطی است که به کمک آن می‌توان یک متغیر پاسخ Y را بر مبنای چند متغیر توضیحی $x_1, \dots, x_p$ توصیف کرده و از آن برای پیش‌بینی مقادیر دیگر متغیر Y استفاده کرد. یکی از حالت‌های خاص مدل (۲-۱) مدل رگرسیون خطی ساده به فرم

$$Y = X\beta + \varepsilon \qquad (۴-۲)$$

است که در آن Y بردار $n \times 1$ متغیر پاسخ، X ماتریس $n \times p$ از متغیرهای توضیحی، β بردار p بعدی پارامترهای براوردپذیر و ε بردار n بعدی عبارت خطا است. براورد بردار پارامترهای β بر مبنای یک نمونه‌ی n تایی از مقادیر مشاهده شده‌ی y_i، $i = 1, \dots, n$ و متغیرهای توضیحی مربوط به آن‌ها انجام می‌شود. برای هر مشاهده‌ی خاص، مثلاً مشاهده‌ی iام، اگر x_i^T ردیف iام ماتریس X باشد، می‌توان مدل را به‌صورت $y_i = x_i^T\beta + \varepsilon_i$ نوشت که در آن

$$E(\varepsilon_i) = 0, \quad var(\varepsilon_i) = \sigma^2$$

و برای هر دو مشاهده‌ی i و j $(i \neq j)$، $cov(\varepsilon_i, \varepsilon_j) = 0$ می‌باشد. در این حالت امید ریاضی مشاهدات به فرم حاصل‌ضرب ماتریس X در بردار β به شکل $E(Y) = X\beta$ و ماتریس واریانس مشاهدات را به فرم حاصل‌ضرب σ^2 در ماتریس همانی می‌باشد؛ پس $var(Y) = \sigma^2 I_n$.

۲-۳-۱ براورد پارامترها و خواص آن‌ها

قضیه (۲-۲). اگر مدل، به‌صورت مدل (۲-۴) در نظر گرفته شود، آن‌گاه براورد کمترین توان‌های دوم بردار پارامترهای β عبارت است از

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \qquad (۵-۲)$$

اثبات: رنچر و اسکالچ (۲۰۰۷) صفحه ۱۴۲.

این براوردگرها در مقایسه با دیگر براوردها از خواص مطلوب‌تری برخوردارند از جمله این‌که دارای توزیع مجانبی

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \to N_p(0, \sigma^2 (X^T X)^{-1}) \qquad (۶-۲)$$

می‌باشند. در ادامه به برخی دیگر از خواص این براوردگرها اشاره خواهیم کرد (رنچر و اسکالچ، ۲۰۰۷).

قضیه‌ی (۲-۳). اگر $E(Y)=X\beta$ و $var(Y)=\sigma^2 I_n$ باشد، آن‌گاه براوردگرهای کمترین مربعات $\hat{\beta}_j$، $j=1,\dots,p$، در میان تمام براوردگرهای نااریب خطی، دارای مینیمم واریانس هستند.

اثبات: رنچر و اسکالچ (۲۰۰۷) صفحه ۱۴۲.

نتیجه (۲-۱). اگر $E(Y)=X\beta$ و $var(Y)=\sigma^2 I_n$ باشد، بهترین براوردگر خطی نااریب $a^T\beta$ عبارت است از $a^T\hat{\beta}$ که در آن $\hat{\beta}$ در رابطه‌ی (۲-۵) تعریف شده است.

اثبات: رنچر و اسکالچ (۲۰۰۷) صفحه ۱۴۸.

قضیه (۲-۴). اگر $var(Y)=\sigma^2 I_n$ باشد، کوواریانس ماتریس $\hat{\beta}$ برابر

$$var\left(\hat{\beta}\right)=\sigma^2 (X^T X)^{-1} \qquad (۷-۲)$$

خواهد بود.

اثبات: رنچر و اسکالچ (۲۰۰۷) صفحه ۱۴۵.

بردار مقادیر برازش شده عبارت است از:

$$\hat{Y}=X\hat{\beta}=X(X^T X)^{-1}X^T Y=HY. \qquad (۸-۲)$$

در تعریف فوق، H، ماتریس تصویر[۱] و برابر

$$H=X(X^T X)^{-1}X^T \qquad (۹-۲)$$

[۱] Projection matrix

می‌باشد که با استفاده از آن بردار باقی‌مانده‌ها را به شکل

$$e = Y - \hat{Y} = Y - X\hat{\beta} = (I - H)Y \qquad (۱۰ - ۲)$$

تعریف می‌شود و می‌توان از آن به‌عنوان جانشینی برای ε استفاده کرد. با استفاده از رابطه‌ی (۲-۱۰) مقدار باقی‌مانده‌ی iام برابر $e_i = y_i - \hat{y}_i$ خواهد بود. اگر طبق رابطه‌ی (۲-۴) به جای Y، از مقدار معادل آن استفاده شود، رابطه‌ی میان e و ε به‌دست می‌آید:

$$e = (I - H)(X\beta + \varepsilon)$$
$$= (I - H)\varepsilon.$$

این رابطه صراحتاً بیان می‌کند که رابطه‌ی میان e و ε تنها به ماتریس H بستگی دارد و با استفاده از آن، می‌توان e_i را به فرم اسکالر $e_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j$ نوشت که

$$h_{ij} = x_i^T (X^T X)^{-1} x_j \qquad (۱۱ - ۲)$$

عضو (i,j)ام ماتریس H می‌باشد و در آن x_i^T سطر iام ماتریس X است. اگر h_{ij}ها به اندازه‌ی کافی کوچک باشند، e جانشین مناسبی برای ε خواهد بود (رنچر و اسکالچ[۱]، ۲۰۰۷).

درک بهتر رابطه‌ی میان e و ε مستلزم مطالعه‌ی دقیق‌تر پیرامون ماتریس H می‌باشد. به دلیل اهمیت بردار باقی‌مانده‌ها و ماتریس تصویر، در فصل آینده به تفصیل درباره‌ی آن‌ها بحث خواهیم کرد.

از سوی دیگر آماره‌ی مجموع مربعات باقی‌مانده‌ها به صورت

$$s^2 = e^T e \qquad (۱۲ - ۲)$$

تعریف شود. می‌توان از

$$\hat{\sigma}^2 = \frac{s^2}{n - p} \qquad (۱۳ - ۲)$$

[۱] Rencher & Schaalje