دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

تجزیه و تحلیل رگرسیون مولفه‌های اصلی

پایان‌نامه کارشناسی ارشد آمار اقتصادی و اجتماعی

بتول ذکیا

اساتید راهنمای پایان‌نامه

دکتر علی زینل همدانی
دکتر سروش علیمرادی

۱۳۸۵

جلسه دفاع از پایان‌نامه کارشناسی ارشد

تجزیه و تحلیل رگرسیون مولفه‌های اصلی

سخنران: بتول ذکیا

زمان: چهارشنبه ۸۵/۱۰/۱۳ ساعت ۱۶/۵

مکان: سالن کنفرانس دانشکده علوم ریاضی

هیئت داوران

۱– دکتر علی زینل همدانی

۲– دکتر سروش علیمرادی

۳– دکتر حمید پزشک (گروه آمار دانشگاه تهران)

۴– دکتر سعید پولاد ساز

چکیده:

رگرسیون چندگانه با متغیرهای توضیحی به‌هم وابسته در دامنه وسیعی از علوم از جمله علوم اجتماعی، اقتصادی، مهندسی، کشاورزی و پزشکی کاربرد دارد. در مدل رگرسیون چندگانه اگر متغیرهای توضیحی وابسته باشند، آنگاه برآورد ضرایب رگرسیونی بسیار نادقیق خواهد بود، از طرفی افزایش تعداد متغیرهای توضیح دهنده موجب وجود همبستگی بین متغیرها می‌گردد و لذا محققین برای رفع این مشکل از مولفه‌های اصلی که هدف اصلی آنها کاهش ابعاد متغیرها و تولید متغیرهای جدید توضیح‌دهنده مستقل می‌باشد، استفاده می‌کنند. در این پایان نامه با مروری اجمالی بر روشهای تجزیه و تحلیل رگرسیون چندگانه و مولفه‌های اصلی روش انتخاب مناسب‌ترین مولفه‌های اصلی برای تجزیه و تحلیل رگرسیون چندگانه مورد بحث و بررسی قرار می‌گیرد و در نهایت با استفاده از شبیه‌سازی روشهای مورد بحث مورد ارزیابی قرار می‌گیرند.

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد آمار اقتصادی و اجتماعی خانم بتول ذکیا

تحت عنوان

تجزیه و تحلیل رگرسیون مولفه‌های اصلی

در تاریخ ۸۵/۱۰/۱۳ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهائی قرار گرفت.

۱– استاد راهنمای پایان‌نامه	دکتر علی زینل همدانی
۲– استاد راهنمای پایان‌نامه	دکتر سروش علیمرادی
۳– استاد داور ۱	دکتر حمید پزشک (گروه آمار دانشگاه تهران)
۴– استاد داور ۲	دکتر سعید پولاد ساز

سرپرست تحصیلات تکمیلی دانشکده دکتر رسول نصر اصفهانی

سپاس بر آستان معبودی می‌ستایم که حکمت و دست پر توان اراده‌اش مرا در اقیانوس ژرف و پر تلاطم زندگی، بی‌یار و ناامید نگذاشت و شالوده علم و دانش را در وجودم سرشت و جانم را تشنه آموختن آفرید. او را سپاس می‌گویم که زیادت خواه نعمت او و گردن نهاده عزت اویم.

زلال‌ترین سپاس‌ها را تقدیم می‌کنم به:

پدر و مادر عزیزم، آن‌ها که در شکیب لحظه‌ها بر تازیانه تیره مکتوب، همچون نور از پس لطافت الهام‌گون ابر صبر و تلاش را به هم‌آوازی امید بر لوح جانم نقش زدند.

و به یگانه خواهر مهربان و برادران خوبم، محمد و وحید.

از جناب آقای دکتر علی همدانی و جناب آقای دکتر سروش علیمرادی، اساتید بزرگوار و ارجمندم که افق روشنی را در برابر اندیشه‌هایم به تصویر کشیدند، بسیار سپاسگزارم. به عجز کلمات از واگویه این سپاس تنها همین بس که اگر همکاری و مساعدت‌های ایشان نبود، گذر از این عرصه‌ی علمی میسر نمی‌نمود.

از جناب آقای دکتر حمید پزشک و جناب آقای دکتر سعید پولادساز که زحمت داوری و بازخوانی این رساله را بر خود هموار نمودند سپاسگزارم.

همچنین از تمام اساتید بزرگواری که در دوران تحصیل بر کرس دانشجویی ایشان نشسته‌ام بسیار سپاسگزارم.

از همه دوستان خوبم که آشنایی و همراهی‌شان فرصتی تکرار ناشدنی بود و از هر یک از آنها به فراخور حال نکات زیادی آموختم و یادشان همیشه همراه من خواهد بود بسیار سپاسگزارم و برای هر یک از آنها آرزوی کامیابی و پیروزی دارم.

در پایان نیز از همکاری تمامی بخش‌های دانشکده علوم ریاضی به ویژه سرکار خانم دلیلی و سرکار خانم زابلیان و سرکار خانم صدر عاملی کمال تشکر را دارم.

فهرست مطالب

چکیده:

رگرسیون چندگانه با متغیرهای توضیحی به‌هم وابسته در دامنه وسیعی از علوم از جمله علوم اجتماعی، اقتصادی، مهندسی، کشاورزی و پزشکی کاربرد دارد. در مدل رگرسیون چندگانه اگر متغیرهای توضیحی وابسته باشند، آنگاه برآورد ضرایب رگرسیونی بسیار نادقیق خواهد بود، از طرفی افزایش تعداد متغیرهای توضیح دهنده موجب وجود همبستگی بین متغیرها می‌گردد و لذا محققین برای رفع این مشکل از مولفه‌های اصلی که هدف اصلی آنها کاهش ابعاد متغیرها و تولید متغیرهای جدید توضیح‌دهنده مستقل می‌باشد، استفاده می‌کنند. در این پایان نامه با مروری اجمالی بر روشهای تجزیه و تحلیل رگرسیون چندگانه و مولفه‌های اصلی روش انتخاب مناسب‌ترین مولفه‌های اصلی برای تجزیه و تحلیل رگرسیون چندگانه مورد بحث و بررسی قرار می‌گیرد و در نهایت با استفاده از شبیه‌سازی روشهای مورد بحث مورد ارزیابی قرار می‌گیرند.

فصل ۱

مقدمات و تعاریف اساسی

۱—۱ مقدمه

مشکل دنیای امروز نبود داده و اطلاعات کافی برای تصمیم‌گیریهای علمی نیست. بلکه محققان در بیشتر زمینه‌های مطالعاتی با سیلانی از داده‌های خام مواجه هستند که برای ارائه تحلیل‌های مفید و کارآمد نیازمند روشی مناسب برای استخراج اطلاعات می‌باشند.

وقتی تعداد متغیرها زیاد است، مشکلات عمده‌ای در تحلیل داده‌ها وجود دارد. اولاً برخی از متغیرها ممکن است بطور کلی با هدف تحقیق ارتباط نداشته باشند و شاید نگاه داشتن این متغیرها اثر خالص متغیرهای دیگر را تحت تاثیر قرار دهد. ثانیاً باعث تعداد زیاد پارامتر در مساله می‌شود و در نتیجه تابع هدف برای برآورد پارامترها پیچیده‌تر می‌شود. از طرفی در بسیاری از علوم کاربردی کمتر حالتی پیش می آید که

متغیرهای مستقل، متعامد باشند، بلکه در بعضی از مواقع یک ارتباط تقریباً خطی بین متغیرهای مستقل وجود دارد. در این حالت برآورد مدل از طریق حداقل مربعات معمولی دارای خطای زیادی می‌باشد.
بنابراین برای تقلیل در بعد مساله ناگزیریم تعدادی از این متغیرها را کنار بگذاریم. یک راه‌حل، حذف یکی از زوج متغیرهای توضیحی است که شدیداً همبسته‌اند. در اینصورت مشکلات هم‌خطی تا حدودی رفع می‌شود. روشهای مختلفی برای انتخاب متغیر در رگرسیون چندگانه وجود دارد که استفاده از آنها برای رسیدن به حداکثر نیکویی برازش مفید است.
استفاده از رگرسیون مولفه‌های اصلی یک راه‌حل برای این مشکلات است. بدین‌ترتیب که مولفه‌های اصلی متغیرها را محاسبه کرده، سپس متغیر پاسخ را روی زیرمجموعه‌ای از این مولفه‌ها برازش می‌دهیم. با استفاده از این روش تعداد متغیرهای مستقل به‌صورت قابل ملاحظه‌ای کاهش می‌یابد و از آنجاییکه مولفه‌های اصلی دوبدو ناهمبسته‌اند مشکلات هم‌خطی چندگانه برای مدل بوجود نمی‌آید.
بنابراین هرگاه هم‌خطی چندگانه وجود دارد، رگرسیون مولفه‌های اصلی بعنوان یک ابزار پرتوان برای تجزیه و تحلیل حجم بالای داده‌ها و پایگاه‌های داده‌ای با ابعاد زیاد، بکار می‌رود. در زیر مروری بر مطالعات انجام شده در این زمینه خواهیم داشت.

۱—۲ مروری بر مطالعات انجام شده

تاریخچه پیدایش رگرسیون مولفه‌های اصلی، بر خلاف مولفه‌های اصلی چندان گسترده نیست. آنچه در این بخش بعنوان تاریخچه ارائه می‌گردد، خلاصه‌ای از کارهایی است که آماردانان بر روی این مبحث انجام داده‌اند.
حدوداً ۴۱ سال پیش ویلیام میسی [۱] [۱۹] در سال (۱۹۶۵)، رگرسیون مولفه‌های اصلی را برای بررسی ارتباط بین سطح تحصیلات و میزان درآمد بر روی داده‌های سرشماری ۱۹۵۰ در شهر شیکاگو بکار برد.
مارکوارت (۱۹۷۰) [۱۷] در مقاله‌ای تحت عنوان « معکوس تعمیم یافته، رگرسیون مرزی، برآورد خطی و غیر خطی اریب» یک شرط کافی برای برتری برآوردگر مولفه‌های اصلی نسبت به برآوردگر حداقل مربعات معمولی ارائه داد. در اوائل ۱۹۷۷ ریچارد، گونست و میسون [۸] در یک مطالعه شبیه‌سازی رگرسیون مولفه‌های اصلی را با رگرسیون حداقل مربعات معمولی، رگرسیون ریشه ویژه و رگرسیون انقباضی و رگرسیون مرزی مورد مقایسه قرار دادند. دمپستر، اسکاتزف و ورمیوت [۲] نیز در همین سال در مطالعه شبیه‌سازی دیگری رگرسیون مولفه‌های اصلی و سایر رگرسیون‌های اریب را برای ۱۶۰ مدل مختلف

[۱] William Massy

مهاجرت مقایسه کردند.

در سال ۱۹۸۲ جولایف [۲] [۱۵] ، کتابی تحت عنوان تجزیه و تحلیل مولفه‌های اصلی را منتشر کرد و ویرایش بعدی آن در سال ۱۹۸٦ انتشار یافت. جولایف در این کتاب رگرسیون مولفه‌های اصلی و روشهای انتخاب مولفه‌ها در رگرسیون مولفه‌های اصلی را مورد بحث قرار داده‌است. فرانک و فریدمن (۱۹۹۳)[۳] [٦] در مطالعه شبیه‌سازی دیگری رگرسیون مولفه‌های اصلی و رگرسیون حداقل مربعات جزئی و رگرسیون مرزی را مورد مقایسه قرار دادند.

هادی و لینگ (۱۹۹۸) [۱۰] در مقاله‌ای تحت عنوان «چند اخطار در استفاده از رگرسیون مولفه‌های اصلی» ضعفها و ویژگی‌های رگرسیون مولفه‌های اصلی را بر روی مجموعه داده‌های لانگلی و هالد بررسی کردند.

در سال ۲۰۰۱ فیلزموسر [۴] رگرسیون مولفه‌های اصلی نیرومند را معرفی کرد و سپس این روش را در روی مجموعه داده‌های زمین شناسی اجرا کرد. فیلزموسر و کراکس (۲۰۰۳) [۴] [۵] یک روش جدید برای کاهش ابعاد متغیرهای توضیحی و ساخت مولفه‌ها در رگرسیون مولفه‌های اصلی ارائه کردند. در این سال همچنین هانگ و نتلتون [۵] [۱۲] یک دیدگاه استنباطی برای انتخاب مولفه‌ها در رگرسیون مولفه‌های اصلی بیان کردند. آنها همچنین برآوردگرهای متعددی برای رگرسیون مولفه‌های اصلی معرفی کردند و در یک مطالعه شبیه‌سازی MSE آنها را مقایسه و بهترین برآوردگرها را انتخاب کردند.

۱—۳ تعاریف و مفاهیم

مقادیر ویژه و بردارهای ویژه[٦]

فرض کنید $A_{n\times n}$، ماتریس تبدیل بردار X باشد که مضربی از X را نتیجه می‌دهد، یعنی اگر $X_{n\times 1} \in E_n$ آنگاه

$$AX = lX$$

بنابراین اگر X وجود داشته باشد، معادله $(A - lI)X = 0$ را نتیجه می گیریم که یک جواب بدیهی آن $X = 0$ است.

---

[۲] Jollife,I.T

[۳] E.Frank and H.Friedman

[۴] Filzmoser,P. and Croux,C.

[۵] Hwang, J.G and Netteleton,D.

[٦] Eigenvalues and Eigenvectors

تعریف: فرض کنید A یک ماتریس مربع $n \times n$ و X یک بردار غیر صفر است. l را مقدار ویژه ماتریس A گوییم اگر $AX = lX$ و بردار X متناظر با l را بردار ویژه ماتریس A نامیم. برای یافتن lها می‌بایست معادله $|A - lI| = ۰$ را حل کرد که حاصل آن کثیرالجمله‌ای از رتبه n و برحسب l است. معادله حاصل را معادله ویژه می‌نامیم و بصورت زیر نمایش می‌دهیم:

$$a_n l^n + a_{n-۱} l^{n-۱} + \cdots + a_۱ l + a_۰ = ۰$$

پس از حل معادله ویژه، n مقدار ویژه بدست می آید که با استفاده از معادله $AX = lX$ بردارهای ویژه متناظر مقادیر ویژه تعیین می‌شوند.[۲۲]

بردار متعامد [۷] :

دو بردار u و v متعامد نامیده می‌شوند اگر حاصلضرب داخلی آنها صفر باشد. یعنی:

$$u.v = ۰$$

در فضای سه بعدی بردارهای متعامد دوبه‌دو بر هم عمودند.[۲۲]

بردار یکامتعامد [۸] :

زیرمجموعه $\{v_۱, ..., v_k\}$ از فضای برداری V یکامتعامد نامیده می‌شود اگر حاصلضرب داخلی دو بردار v_i و v_j برابر صفر بوده و طول هر یک از بردارها برابر یک باشد. یعنی:

$$< v_i, v_j >= ۰ \qquad i \neq j$$

$$< v_j, v_j >= ۱$$

بعبارت دیگر بردارها دوبه‌دو برهم عمودند.[۲۲]

پایه [۹] :

مجموعه $B = \{X_۱, X_۲, ..., X_n\}$ شامل n بردار غیر صفر در R^n که دوبه‌دو برهم عمودند را یک پایه برای R^n می‌نامند.[۲۲]

پایه یکامتعامد [۱۰] :

پایه $B = \{X_۱, X_۲, ..., X_n\}$ در R^n پایه متعامد نامیده می‌شود اگر عناصر B دوبه‌دو متعامد باشند. یعنی:

$$X_i.X_j = ۰ \qquad i \neq j$$

---

[۷] Orthogonal Vector
[۸] Orthonormal Vector
[۹] basis
[۱۰] Orthonormal basis

همچنین این پایه، پایه یکامتعامد نامیده می‌شود اگر شرط زیر را نیز داشته باشد.

$$X_i.X_i = ۱ \qquad \forall i$$

بنابراین پایه یکامتعامد، پایه‌ای است که بردارهای آن طول واحد دارند و دوبه‌دو بر هم عمودند.[۲۲]

تجزیه مقدار منفرد X [۱۱] :

برای هر ماتریس X با ابعاد $n \times p$ می‌توان تجزیه زیر را بدست آورد:

$$X = UDV'$$

که $U'U = V'V = I$ و $D_{p\times p}$ ماتریس قطری با عناصر $\mu_۱, \cdots, \mu_p$ است، μ_iها مقادیر منفرد X می‌باشند و $U_{n\times p}$ دارای ستونهای متعامد و $V_{p\times p}$ دارای سطرها و ستونهای متعامد است. رابطه $X = UDV'$ نشان می‌دهد که ستونهای X ترکیب خطی از ستونهای U و سطرهای X ترکیب خطی از ستونهای V هستند. ارتباط بین مقادیر ویژه و مقادیر منفرد بسادگی قابل دسترسی است. برای این منظور می‌توان نوشت:

$$X'X = VDU'UDV' = VD^۲V'$$

با ضرب طرفین $X'X = VD^۲V'$ در V نتیجه خواهیم گرفت:

$$X'XV = VD^۲$$

اگر فرض کنیم $A = X'X$، از مقایسه دو رابطه $AV = VL$ و $AV = VD^۲$ نتیجه می‌شود که مقادیر منفرد ریشه دوم مقادیر ویژه‌اند. بنابراین هر یک از این دو در اختیار باشند می‌توان دیگری را بسادگی بدست آورد.[۲۴]

تجزیه طیفی X [۱۲] :

فرض کنید ماتریس $C = (x_۱, x_۲, ..., x_n)$، شامل بردارهای ویژه یکامتعامد ماتریس متقارن A باشد، بنابراین C متعامد است و $CC' = I$.

در این حالت $A = CDC'$، که D یک ماتریس قطری از مقادیر ویژه A است. این روش تجزیه طیفی نامیده می‌شود.[۲۴]

بدشرطی یا شرایط بیمارگونه [۱۳] :

وقتی بین ستون‌های ماتریس X وابستگی واقعی وجود دارد، یعنی وقتی یک (یا چند) ستون را می‌توان

[۱۱] Singular Value Decomposition
[۱۲] Spectral Decomposition
[۱۳] Illconditioning

دقیقاً به‌صورت یک ترکیب خطی (با ضرایب عددی مختلف) از سایر ستون‌ها نوشت، در این‌صورت $det(X'X) = ۰$، این پیشامد را بدشرطی یا شرایط بیمارگونه می‌نامند. به این پیشامد به دو طریق می‌توان فکر کرد یا مدل ورای تشخیص است، یعنی پارامترهایی بیشتر از آنچه برای بیان داده‌ها لازم است در آن منظور شده یا داده‌ها برای برآورد مدلی که در نظر گرفته شده اند کافی نیستند.[۳]

فصل ۲

رگرسیون چندگانه و مولفه‌های اصلی

۲–۱ مقایسه‌های عملکرد

هدف از تحلیل رگرسیونی، بررسی ارتباط بین متغیر پاسخ Y و متغیرهای توضیحی $X_1, X_2, \cdots, X_{p-1}$ است. می‌خواهیم بر اساس متغیرهای توضیحی، Y را پیشگویی کنیم. در مقابل ممکن است بخواهیم از مدل برازش یافته، برای فهم بهتر چگونگی ارتباط متغیرهای خاص با متغیر پاسخ استفاده کنیم. در هر دو حالت به مدلی مناسب برای برازش به داده‌ها نیاز داریم. نخست مدل را تخمین می‌زنیم سپس با تغییرات لازم آن را بهبود می‌بخشیم. در بهبود بخشیدن به مدل ممکن است بعضی از متغیرها از مدل خارج شوند، تصمیم گیری در این رابطه که کدام متغیر در مدل باقی بماند و کدام متغیر از مدل خارج شود، اهمیت خاص دارد. فرض می‌کنیم k متغیر توضیحی، که k عدد بزرگی است فراهم است و باید برای انتخاب

متغیرها برای ارائه مدل خوب تصمیم بگیریم. شاید اگر هدف اصلی پیش‌بینی باشد احساس کنیم که باید از تمامی اطلاعات موجود استفاده کنیم و اطلاعی را از دست ندهیم. در بسیاری از موارد مطالعاتی ممکن است گاهی k از n یعنی تعداد مشاهدات بیشتر باشد، در این حالت به مشکل برخورده و استفاده از بخشی از متغیرها مناسبتر است. بیشتر اوقات در انتخاب مدل و پیش‌بینی فرض می‌کنیم که مدل واقعی به شکل زیر است:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

در این حالت باید به شناسایی βهایی بپردازیم که با متغیر پاسخ رابطه معنی‌داری ندارند. این صورت کلی انتخاب مدل است که البته راههای بسیاری مانند حداقل مربعات، رگرسیون مولفه‌های اصلی و ... را پیش‌رو خواهیم داشت.

۲–۲ رگرسیون چندگانه

۲–۲–۱ مدل و مفروضات

مدل رگرسیون خطی چندگانه تعمیمی از مدل رگرسیون خطی ساده است. فرض می‌کنیم p متغیر توضیحی $X_1, X_2, \cdots, X_p$ با متغیر وابسته Y ارتباط دارند. با انتخاب یک نمونه nتایی یک ماتریس به‌صورت $[x_{i1}, x_{i2}, \cdots, x_{ip}, y_i] \quad i = 1, 2, \cdots, n$ بدست می‌آید که برای برآورد پارامترهای خطی زیر مورد استفاده قرار می‌گیرد.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + u_i \qquad i = 1, 2, \cdots, n$$

فرض‌های این مدل بشرح زیر است:

۱ – u_iها مقادیر تصادفی غیر قابل مشاهده خطا هستند که مستقل بوده و دارای توزیع نرمال با میانگین صفر و واریانس σ_u^2 هستند.

۲ – توزیع خطا از توزیع توام $X_1, X_2, \cdots, X_p$ مستقل است. بنابراین

$$E(Y|x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$V(Y|x_1, x_2, \cdots, x_p) = \sigma^2_{Y|X_1, X_2, \cdots, X_p} = \sigma_u^2$$

۳ – پارامترهای $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ مقادیری ثابتند.

از آنجاییکه مشاهدات $y_1, y_2, \cdots, y_n$ نمونه‌ای تصادفی‌اند که دوبه‌دو مستقلند، لذا مقادیر خطا نیز دوبه‌دو مستقلند.

مدل رگرسیون فوق را می‌توان بصورت زیر نوشت:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + u_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + u_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + u_n \end{aligned} \qquad (۱.۲)$$

می‌توان معادلات را به‌صورت ماتریسی به‌فرم $Y = X\beta + u$ نوشت که در آن:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad ; \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad ; \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

۲–۲–۲ برآورد پارامترها

برای برآورد پارامترهای مدل از روش حداقل مربعات خطا استفاده می‌کنیم.

$$\begin{aligned} \epsilon^2 = \sum_{i=1}^{n} u_i^2 &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2 \\ &= (Y - X\beta)'(Y - X\beta) \end{aligned} \qquad (۲.۲)$$

با کمینه نمودن توان دوم خطا به معادلات نرمال زیر می‌رسیم:

$$X'Y = X'X\hat{\beta}$$

با حل این معادله، برآورد پارامترها به‌صورت زیر بدست می آیند:

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

برای به‌دست آوردن $\hat{\beta}$ ماتریس $(X'X)^{-1}$ باید وجود داشته باشد، بعبارت دیگر $(X'X)$ باید پر رتبه باشد. ولی شرط لازم و کافی برای پر رتبه بودن ماتریس $(X'X)$ این است که ستون‌های ماتریس X مستقل خطی باشد. بزرگ بودن ضریب همبستگی بین متغیرها شرط تناقض برای پر رتبه بودن $(X'X)$ است که این خود برازش دادن مدل را بر داده‌ها با مشکل روبرو می‌کند. این مساله را با عنوان هم‌خطی چندگانه مطرح می‌کنیم که در بخشهای آتی به آن خواهیم پرداخت.

هنگامی‌که یک مدل رگرسیون مورد استفاده قرار می‌گیرد، یکی از اولین سوالات آن است که چقدر مدل مورد نظر مناسب است، شاخص‌های متفاوتی برای تشخیص مدل مناسب استفاده می‌شود که از آن جمله می‌توان به R^2، واریانس مدل و ... اشاره نمود.

هدف از یک مدل رگرسیون خطی چندگانه آن است که مدل ساده‌ای تشکیل شود که به‌خوبی پیش‌بینی را انجام دهد. هنگامی‌که یک متغیر نامربوط به مدل وارد می‌شود خطای معیار آن بدون آنکه پیش‌بینی را بهبود بخشد، افزایش می‌یابد. از طرف دیگر اگر متغیرهایی که پیش‌بینی کننده‌های مهمی هستند از مدل کنار گذاشته شوند، مدل دچار تورش می‌شود و نماینده مدل واقعی نخواهد بود.[۱۴]

۲—۲—۳ روشهای انتخاب متغیر

با استفاده از یک دسته متغیر مستقل مشخص می‌توان مدلهای متفاوتی ساخت. هر چه تعداد متغیرهای مستقل افزایش یابد، تعداد مدلهای ممکن نیز افزایش پیدا می‌کند. هرچند با تعداد کم متغیرهای مستقل می‌توان تمام مدلهای ممکن را بررسی کرد. اما روشهای مختلفی وجود دارد که به محاسبات متعددی نیاز ندارد، این روشها به‌صورت متوالی متغیرهایی را به مدل اضافه یا کم می‌کنند. اگر بخواهیم زیرمجموعه‌ای از متغیرها را انتخاب کنیم بنا به شرایط دو روش زیر را داریم.

تمام رگرسیونهای ممکن

از این روش زمانی استفاده می‌شود که تعداد متغیرهای توضیحی که می‌خواهیم وارد مدل کنیم، کم باشد. می‌توان با تعریف معیار نیکویی مدل، این معیار را برای هر زیرمجموعه محاسبه و مقدار بهین آن را انتخاب کرد. این معیار می‌تواند بر پایه اندازه نیکویی برازش یا برآورد خطای پیش‌بینی، یا برآورد تعداد ضرایب غیر صفر، برآورد برخی اندازه‌های فاصله بین مدل بر اساس زیرمجموعه انتخابی و مدل واقعی باشد.[۳]

استفاده از آزمون فرض های دنباله‌ای برای تشخیص βهای غیر صفر

این روش برای انتخاب تعدادی از متغیرهای توضیحی وقتی که تعداد آنها زیاد است کاربرد دارد. از جمله این تکنیک‌ها می‌توان انتخاب پسرو، پیشرو و گام به گام را نام برد. این روشها از نظر محاسباتی ساده‌تر هستند اما ضمانتی برای بهینه بودن مدل نهایی بر اساس معیارهای مورد بحث را ندارند. (برای مطالعه بیشتر به [۳] رجوع کنید.) یک رهیافت متقابل، استفاده از روش‌هایی است که از تمام متغیرها برای برازش استفاده می‌کند، اما برآورد اریب به‌دست می‌دهد. مانند رگرسیون مرزی، انقباضی و رگرسیون مولفه‌های اصلی که در فصول بعدی به تفصیل مورد بررسی قرار می‌گیرند.

۲–۳ مولفه‌های اصلی

در بسیاری از مباحث علمی، در زمینه‌های پزشکی، مدیریت، جامعه‌شناسی و اقتصاد مطالعات بر روی تعداد زیادی از متغیرها انجام می‌پذیرد که همبستگی متقابلی بین این متغیرها وجود دارد و به‌دلیل همین همبستگی در برخی از تجزیه و تحلیل‌ها لازم است تعداد متغیرها را با از دست دادن حداقل اطلاعات کاهش داد. یکی از روشهایی که برای کاهش بعد وجود دارد، روش تجزیه و تحلیل مولفه‌های اصلی است که هتلینگ (۱۹۵۷) ارائه داده است. با توجه به اهمیت استفاده از مولفه‌های اصلی و از آنجاییکه مولفه‌های اصلی ترکیب خطی از تمام متغیرها هستند، ممکن است به آسانی تعبیر و تفسیر نشوند. برای رفع این مشکل، انتخاب یک زیرمجموعه از متغیرهای اصلی که تقریباً همان اطلاعات کلی را داشته باشند راه‌حلی عملی است. بنابراین در یک تحلیل چند متغیره می‌توان مولفه‌های اصلی را بدست آورده و سپس بوسیله معیارهایی، بعضی از متغیرها را در مولفه‌های اصلی حذف نمود. البته این انتخاب متغیر در برخی از روشها بطور برعکس انجام می‌پذیرد. یعنی ممکن است ابتدا زیرمجموعه‌ای از متغیرهای اصلی را انتخاب کرده و سپس تحلیل مولفه‌های اصلی را روی این زیرمجموعه انجام دهیم.

تحلیل مولفه‌های اصلی[۱] به تبیین ساختار واریانس–کواریانس به کمک چند ترکیب خطی از متغیرهای اصلی می‌پردازد. اهداف کلی آن عبارتند از: (۱) کاهش حجم داده‌ها و (۲) تعبیر و تفسیر آنها.

اگر چه برای مطالعه تغییرپذیری کل سیستم، p مولفه لازم است، ولی اغلب این تغییرپذیری را می‌توان با تعداد کمتر مثلاً k مولفه اصلی بیان نمود. در این صورت میزان اطلاعی که در k مولفه وجود دارد (تقریباً) مانند میزان اطلاع در p متغیر اولیه است. بنابراین k مولفه اصلی را می‌توان به جای p متغیر اولیه به‌کار برد و مجموعه داده‌های اولیه که شامل n اندازه روی p متغیر است به مجموعه‌ای از داده‌های شامل n اندازه

[۱] Principle Component