دانشکده علوم ریاضی

پایان‌نامه جهت اخذ درجه کارشناسی ارشد

آمار ریاضی

رگرسیون حداقل مربعات جزئی

استاد راهنما

دکتر مجید سرمد

استاد مشاور

دکتر ناصر رضا ارقامی

نگارش

امیر شاهینی

شهریور ۱۳۸۸

قدردانی

در اینجا دوست دارم از همه تشکر کنم

چکیده

در تحقیقات مختلف با مسائلی سر و کار داریم که در آنها با استفاده از مجموعه‌ای از متغیرهای مستقل به پیشگویی رفتار مجموعه‌ای دیگر از متغیرها می‌پردازیم. یکی از روش‌های آماری که کاربرد وسیعی در اینگونه مسائل دارد رگرسیون چندگانه می‌باشد. اما هنگامی که تعداد متغیرهای مستقل بسیار زیاد است یا حجم نمونه کم است مسأله هم‌خطی چندگانه رخ می‌دهد و در نتیجه رگرسیون حداقل مربعات معمولی به ایجاد برآوردهای ناپایداری از ضرایب رگرسیونی می‌انجامد. راه‌حلی که معمولاً برای رفع مسأله هم‌خطی پیشنهاد می‌شود حذف متغیرهایی است که باعث ایجاد هم‌خطی شده‌اند. اما یکی از اشکالات حذف متغیر این است که در این حالت از روابط موجود بین متغیرهای توضیحی چشم‌پوشی می‌شود. راه‌حل دیگر برای رفع مسأله هم‌خطی استفاده از روش‌های رگرسیونی اریب است که با استفاده از ترکیبات خطی خاص متغیرهای توضیحی مسأله هم‌خطی را رفع می‌کنند. رگرسیون حداقل مربعات جزئی یکی از روش‌های چند متغیره است که در هنگام بروز هم‌خطی بین متغیرهای توضیحی مورد استفاده قرار می‌گیرد و با استفاده از ترکیبات خطی خاص متغیرهای توضیحی ،که مؤلفه نامیده می‌شوند، برآوردگرهایی ایجاد می‌کند که پایداری بیشتری نسبت به برآوردگرهای حاصل از روش حداقل مربعات معمولی دارد.

واژه‌های کلیدی: *اعتبارسنجی متقابل، بردار ویژه، پیشگویی، رگرسیون حداقل مربعات جزئی، رگرسیون مؤلفه‌های اصلی، محموله، مقدار ویژه، مؤلفه، هم‌خطی چندگانه*

پیش‌گفتار

در این پایان‌نامه یکی از روش‌های رگرسیونی اریب به نام **رگرسیون حداقل مربعات جزئی** که به اختصار **PLS** نامیده می‌شود معرفی خواهد شد. این روش بیشتر در مواردی که بین متغیرهای توضیحی هم‌خطی شدید وجود دارد و در نتیجه رگرسیون حداقل مربعات معمولی برآوردهای ناپایداری از ضرایب رگرسیونی ایجاد می‌کند مورد استفاده قرار می‌گیرد.

فصل اول شامل دو بخش کلی است. در بخش اول آن بعضی از تعاریف مربوط به جبر ماتریس‌ها آمده و در بخش دوم نیز مقدماتی از رگرسیون چندگانه ارائه شده است. در پایان این فصل روش تحلیل مؤلفه‌های اصلی به طور کامل تشریح خواهد شد.

در فصل دوم، مسأله هم‌خطی چندگانه در رگرسیون مورد بررسی قرار می‌گیرد. در این بررسی علل ایجاد مسأله هم‌خطی، روش‌های مختلف تشخیص این مسأله و راه‌های رفع آن معرفی می‌شود. سپس دو روش رگرسیونی تحت عنوان **رگرسیون ستیغی** و **رگرسیون مؤلفه‌های اصلی** به عنوان راه‌حل‌هایی برای برخورد با مسأله هم‌خطی معرفی می‌شوند. در پایان این فصل نیز کلیه مطالب گفته شده در مورد رگرسیون مؤلفه‌های اصلی در قالب یک مثال تشریح می‌شود.

در فصل سوم، رگرسیون حداقل مربعات جزئی به عنوان راهکار دیگری برای برخورد با هم‌خطی چندگانه معرفی می‌شود. در بخش اول تاریخچه‌ای در مورد PLS ارائه می‌شود و سپس یکی از الگوریتم‌های آن مورد تفسیر و بررسی قرار می‌گیرد. در ادامه با استفاده از یک مثال، رگرسیون PLS با رگرسیون مؤلفه‌های اصلی مقایسه می‌شود. در پایان فصل سوم نیز روش‌های مختلف اعتبارسنجی مدل‌های رگرسیونی تشریح می‌شود.

در فصل چهارم، رگرسیون PLS با استفاده از شبیه‌سازی با چهار روش دیگر مورد مقایسه قرار می‌گیرد

و نتایج حاصل در پایان فصل ارائه می‌گردد. در واقع هدف از این شبیه‌سازی تشخیص وضعیت‌هایی است که PLS عملکرد بهتری نسبت به دیگر روش‌های رگرسیونی دارد. همچنین کلیه محاسبات با استفاده از نرم‌افزار R انجام شده و کدهای مربوطه نیز در پیوست ب آمده است.

فهرست مطالب

۱ تعاریف و مقدمات **۱**

۱.۱ مقدمه ۱

۲.۱ تعاریف جبری ۱

۳.۱ مباحث آماری ۳

۱.۳.۱ رگرسیون چندگانه ۴

۲.۳.۱ تحلیل مؤلفه‌های اصلی ۱۱

۲ هم‌خطی چندگانه و رگرسیون مؤلفه‌های اصلی **۱۴**

۱.۲ مقدمه ۱۴

۲.۲ هم‌خطی چندگانه ۱۴

۱.۲.۲ علل ایجاد هم‌خطی چندگانه ۱۵

۲.۲.۲ آثار هم‌خطی چندگانه ۱۷

۳.۲.۲ ملاک‌های تشخیص هم‌خطی چندگانه ۱۹

۴.۲.۲ روش‌های غلبه بر هم‌خطی چندگانه ۲۴

۳.۲ رگرسیون مؤلفه‌های اصلی ۲۸

۱.۳.۲ راه‌هایی برای انتخاب مؤلفه‌ها در روش **PCR** ۳۵

۲.۳.۲ یک مثال ۳۷

۳.۳.۲ رابطه بین رگرسیون مؤلفه‌های اصلی و رگرسیون ستیغی ۴۲

۴.۳.۲ متراکم سازی داده‌ها با استفاده از **PCR** ۴۳

۵.۳.۲ مشکل اساسی روش **PCR** ۴۴

۳ رگرسیون حداقل مربعات جزئی ۴۵
۱.۳ مقدمه . . . ۴۵
۲.۳ تاریخچه **PLSR** . . . ۴۵
۳.۳ ساختار کلی مدل **PLS** . . . ۴۷
۴.۳ الگوریتم اصلی **PLS** . . . ۴۹
۵.۳ مثال تنباکو . . . ۵۱
۶.۳ * مقایسه **PLS** و **PCR** . . . ۵۳
۷.۳ تفسیر جواب‌ها در **PLS** . . . ۵۷
۸.۳ اعتبارسنجی . . . ۵۹
۱.۸.۳ اعتبارسنجی با استفاده از مجموعه داده‌های اصلی . . . ۶۰
۲.۸.۳ سنجش پیشگویی . . . ۶۰
۳.۸.۳ اعتبارسنجی متقابل . . . ۶۲

۴ *شبیه‌سازی ۶۶
۱.۴ مقدمه . . . ۶۶
۲.۴ مجموعه داده‌های **NIR** کیوی . . . ۶۶
۳.۴ نحوه شبیه‌سازی . . . ۶۷
۱.۳.۴ محاسبه مقادیر پارامتر . . . ۶۸
۲.۳.۴ روش‌های رگرسیونی . . . ۶۸
۳.۳.۴ تابع زیان و روش شبیه‌سازی . . . ۷۰
۴.۴ نتایج شبیه‌سازی . . . ۷۱
۱.۴.۴ مقایسه نتایج با نتایج حاصل از شبیه‌سازی‌های دیگر . . . ۷۲

کتاب‌نامه ۷۷

الف واژه‌نامه ۱

ب کدهای برنامه‌نویسی ۵

لیست تصاویر

۱.۱ انتخاب تعداد بهینه متغیرها برای جلوگیری از رخداد بیش برازش یا کم برازش ۱۱

۱.۲ توزیع نمونه‌ای برآوردگرهای اریب و نااریب ۲۶

۲.۲ متراکم‌سازی داده‌ها با استفاده از PCR ۴۴

۱.۳ $RMSEP$ دو روش PLS و PCR در مقابل تعداد مؤلفه‌ها برای داده‌های تنباکو ۵۴

۲.۳ نمودار مقادیر پیشگویی PCR در مقابل مقادیر واقعی برای داده‌های تنباکو ۵۴

۳.۳ نمودار مقادیر پیشگویی PLS در مقابل مقادیر واقعی برای داده‌های تنباکو ۵۵

۴.۳ $RMSEP$ دو روش PLS و PCR در مقابل تعداد مؤلفه‌ها برای داده‌های کیوی ۵۷

۵.۳ نمودار $MSEP(= RMSEP^۲)$ به عنوان تابعی از تعداد مؤلفه‌ها ۶۲

۱.۴ نمودار میانگین زیان پنج روش رگرسیونی با هشت متغیر توضیحی در مقابل واریانس خطا ۷۴

۲.۴ نمودار میانگین زیان پنج روش رگرسیونی با ۲۰ متغیر توضیحی در مقابل واریانس خطا ۷۵

۳.۴ نمودار میانگین زیان پنج روش رگرسیونی با ۵۰ متغیر توضیحی در مقابل واریانس خطا ۷۶

لیست جداول

۱.۲ مجموعه داده‌های تنباکو ۳۸
۲.۲ عوامل تورم واریانس داده‌های تنباکو ۴۰
۳.۲ مقادیر ویژه ماتریس همبستگی داده‌های تنباکو ۴۰
۴.۲ بردارهای ویژه ماتریس همبستگی داده‌های تنباکو ۴۰
۵.۲ مؤلفه‌های اصلی داده‌های تنباکو ۴۱

۱.۳ محموله‌های وزنی PLS برای داده‌های تنباکو ۵۳
۲.۳ محموله‌های PLS برای داده‌های تنباکو ۵۳
۳.۳ ضرایب رگرسیونی PLS برای داده‌های تنباکو بر اساس متغیرهای X ۵۳
۴.۳ انتخاب تعداد مؤلفه‌های PLS بر اساس معیار $RMSEP$ برای داده‌های تنباکو ۵۳
۵.۳ مؤلفه‌های PLS برای داده‌های تنباکو ۶۴
۶.۳ عوامل تورم واریانس داده‌های کیوی ۶۵
۷.۳ مقادیر ویژه ماتریس همبستگی داده‌های کیوی ۶۵
۸.۳ معیار $RMSEP$ برای روش‌های PLS و PCR در داده‌های کیوی ۶۵

۱.۴ میانگین زیان پیشگویی پنج روش رگرسیونی با هشت متغیر توضیحی و واریانس‌های خطای مختلف . . ۷۴
۲.۴ میانگین زیان پیشگویی پنج روش رگرسیونی با ۲۰ متغیر توضیحی و واریانس‌های خطای مختلف . . . ۷۵
۳.۴ میانگین زیان پیشگویی پنج روش رگرسیونی با ۵۰ متغیر توضیحی و واریانس‌های خطای مختلف . . . ۷۶

فصل ۱

تعاریف و مقدمات

۱.۱ مقدمه

این فصل به یادآوری برخی تعاریف و مقدمات در مباحث جبری و آماری اختصاص داده شده است. کتاب‌های [۱۳]، [۱۵] و [۱۹] عنوان مراجع اصلی مطالب این فصل می‌باشند.

۲.۱ تعاریف جبری

در این بخش برخی از تعاریف مورد نیاز در جبر خطی آورده شده است.

تعریف ۱.۲.۱. فرض کنید $\mathbf{A}$ یک ماتریس مربع $p \times p$ و $\mathbf{I}$ نیز ماتریس همانی از همان مرتبه باشد. اسکالرهای $\lambda_1, \lambda_2, \cdots, \lambda_p$ که در معادله‌ی چند جمله‌ای $|\mathbf{A} - \lambda\mathbf{I}| = 0$ صدق می‌کنند را **مقادیر ویژه** (ریشه‌های مفسر) ماتریس $\mathbf{A}$ می‌نامند.

تعریف ۲.۲.۱. گوییم ماتریس مربع $\mathbf{A}$ دارای یک مقدار ویژه‌ی λ با **بردار ویژه** $\mathbf{a} \neq 0$ است، هرگاه

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{a}$$

شایان ذکر است که بردارهای ویژه یکتا هستند مگر اینکه دو یا چند مقدار ویژه با هم برابر باشند. برای هر ماتریس مربع متقارن $p \times p$، تعداد p زوج مقدار ویژه–بردار ویژه به صورت زیر وجود دارد

$$(\lambda_1, \mathbf{a}_1), (\lambda_2, \mathbf{a}_2), \ldots, (\lambda_p, \mathbf{a}_p)$$

تعریف ۱.۲.۱. بردارهای ویژه را می‌توان طوری انتخاب کرد که بر هم عمود باشند و طول آنها نیز برابر یک باشد. به این کار استانداردسازی و بردارهای حاصل را بردارهای **یکا متعامد** می‌گویند. در این صورت بردارهای استاندارد شده که ما آنها را با $\mathbf{e}_i$ نمایش می‌دهیم در شرایط زیر صدق خواهند کرد

$$\mathbf{e}_i'\mathbf{e}_i = ۱ \qquad i = ۱, ۲, \ldots, p$$
$$\mathbf{e}_i'\mathbf{e}_j = ۰ \qquad i \neq j$$

نکته ۱.۲.۱. *اگر* $\mathbf{A}$ *متقارن باشد مقادیر ویژه‌ی آن حقیقی و بردارهای ویژه‌ی آن متعامد خواهند بود.*

نکته ۱.۲.۱. *هر ماتریس مربعی مانند* $\mathbf{A}$ *را می‌توان به طور یکتا به شکل زیر نمایش داد*

$$\mathbf{A} = \mathbf{\Gamma\Lambda\Gamma}'$$

که در آن $\mathbf{\Gamma}$ *و* $\mathbf{\Lambda}$ *به ترتیب ماتریس‌های متعامد و قطری می‌باشند.*

بنابراین، اگر $\mathbf{\Gamma}$ ماتریسی باشد که ستون‌هایش بردارهای ویژه‌ی $\mathbf{A}$ و $\mathbf{\Lambda}$ نیز ماتریسی شامل مقادیر ویژه‌ی $\mathbf{A}$ باشد آنگاه رابطه‌ی فوق همچنان برقرار خواهد بود.

تعریف ۱.۲.۱. اگر $\mathbf{X}$ یک ماتریس $n \times p$ باشد **تجزیه مقدار- منفرد** آن را به صورت زیر نشان می‌دهیم

$$\mathbf{X} = \mathbf{UD\Gamma}'$$

که در آن $\mathbf{U}$ ماتریس متعامد $n \times p$ است که ستون‌هایش متناظر با بردارهای ویژه مربوط به p مقدار ویژه ناصفر ماتریس $\mathbf{X'X}$ می‌باشد.
$\mathbf{D}$ ماتریس قطری $p \times p$ با اعضای قطری نامنفی μ_j ،برای $j = ۱, ۲, \cdots, p$، می‌باشد که آنها را **مقادیر منفرد** $\mathbf{X}$ می‌نامیم.
$\mathbf{\Gamma}$ نیز ماتریسی است که ستون‌هایش بردارهای ویژه $\mathbf{X'X}$ می‌باشند.

نکته ۱.۲.۱. *اگر بردارهای ویژه و مقادیر ویژه ماتریس* $\mathbf{X'X}$ *را به ترتیب با* $\mathbf{a}_j$ *و* λ_j *،برای* $j = ۱, ۲, \ldots, p$، *نشان دهیم همواره داریم*

$$\mu_j^۲ = \lambda_j \qquad \forall j$$
$$\mathbf{u}_j = \mu_j^{-۱}\mathbf{X}\mathbf{a}_j \quad \forall j$$

که در آن $\mathbf{u}_j$ *ستون* j *ام ماتریس* $\mathbf{U}$ *می‌باشد.*

با توجه به نکته‌ی فوق مشاهده می‌کنیم که تجزیه مقدار–منفرد در ارتباط نزدیکی با مفاهیم مقدار ویژه و بردار ویژه است و می‌توان نوشت

$$\begin{aligned}\mathbf{X}'\mathbf{X} &= (\mathbf{U}\mathbf{D}\boldsymbol{\Gamma}')'(\mathbf{U}\mathbf{D}\boldsymbol{\Gamma}') \\ &= \boldsymbol{\Gamma}\mathbf{D}^{۲}\boldsymbol{\Gamma}' \\ &= \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'\end{aligned}$$

تعریف ۸.۲.۱. مجموع اعضای قطر اصلی یک ماتریس مربعی مانند $\mathbf{A}$ را **اثر** ماتریس $\mathbf{A}$ می‌نامند و با نماد $\mathbf{Tr}$ نمایش می‌دهند.

نکته ۹.۲.۱. *اثر یک ماتریس همواره برابر مجموع مقادیر ویژه آن ماتریس می‌باشد.*

تعریف ۱۰.۲.۱. بردارهای $\mathbf{x}_۱, \mathbf{x}_۲, \ldots, \mathbf{x}_p$ به صورت خطی **وابسته** هستند هرگاه مجموعه‌ای از مقادیر ثابت $t_۱, t_۲, \ldots, t_p$ که همه صفر نیستند وجود داشته باشند به طوری که

$$\sum_{i=۱}^{p} t_j \mathbf{x}_j = ۰ \qquad (۱.۱)$$

نکته ۱۱.۲.۱. *اگر بردارهای $\mathbf{x}_۱, \mathbf{x}_۲, \ldots, \mathbf{x}_p$ وابسته خطی نباشند گفته می‌شود آنها* ***مستقل خطی*** *هستند.*

تعریف ۱۲.۲.۱. ماتریس مربع $\mathbf{A}$ ناتکین نامیده می‌شود هرگاه رتبه‌ی آن برابر تعداد ستون‌هایش (سطرهایش) باشد. در واقع شرط ناتکین بودن به این معنی است که ستون‌های (سطرهای) ماتریس مستقل خطی باشند.

تعریف ۱۳.۲.۱. تجزیه طیفی یک ماتریس متقارن $p \times p$ مانند $\mathbf{A}$ به صورت زیر می‌باشد

$$\mathbf{A} = \sum_{i=۱}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i'$$

که در آن، λ_iها مقادیر ویژه و $\mathbf{e}_i$ ها بردارهای ویژه‌ی استاندارد شده هستند.

۳.۱ مباحث آماری

در این بخش، ابتدا مقدماتی از رگرسیون چندگانه را مورد بررسی قرار می‌دهیم و سپس در مورد یکی از مباحث مهم آماری با عنوان **تحلیل مؤلفه‌های اصلی** که اهمیت فراوانی در مسائل کاربردی دارد صحبت خواهیم کرد.

۱.۳.۱ رگرسیون چندگانه

مدل رگرسیونی که مشتمل بر بیش از یک متغیر توضیحی باشد مدل رگرسیون چندگانه نامیده می‌شود. در حالت کلی متغیر پاسخ Y ممکن است به k متغیر توضیحی بستگی داشته باشد. در این صورت مدل رگرسیون خطی چندگانه با k متغیر توضیحی به صورت زیر نمایش داده می‌شود

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \qquad (۲.۱)$$

که در آن پارامترهای β_j، $j = ۱, ۲, \ldots, k$، ضرایب رگرسیون نامیده می‌شوند. این مدل یک ابرصفحه در فضای k بعدی که توسط متغیرهای توضیحی X_j ساخته شده تشکیل می‌دهد. پارامتر β_j نشان دهنده متوسط تغییرات متغیر پاسخ به ازای یک واحد تغییر در X_j است وقتی متغیرهای توضیحی دیگر ثابت نگه داشته شوند. به همین جهت پارامترهای β_j ضرایب رگرسیون جزئی نامیده می‌شوند. همچنین هر مدل رگرسیون که بر حسب این ضرایب رگرسیونی خطی باشد مدل رگرسیون خطی نامیده می‌شود.

برآوردگر حداقل مربعات ضرایب رگرسیون

فرض می‌کنیم $n > k$ مشاهده در دسترس است و y_i نشان دهنده i امین پاسخ مشاهده شده و x_{ij} نشان دهنده i امین مشاهده روی متغیر توضیحی X_j باشد. همچنین خطاهای تصادفی ε دو به دو ناهمبسته بوده و دارای میانگین صفر و واریانس ثابت σ^2 هستند. در این صورت شکل ماتریسی مدل رگرسیون به صورت زیر است

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

که در آن $\mathbf{Y}$ یک بردار $n \times ۱$، $\mathbf{X}$ یک ماتریس $n \times p$، $\boldsymbol{\beta}$ یک بردار $p \times ۱$ و $\boldsymbol{\epsilon}$ نیز یک بردار $n \times ۱$ از خطاهای تصادفی است به طوری که $p = k + ۱$.

حال بردار برآوردگرهای حداقل مربعات $\hat{\boldsymbol{\beta}}$ را طوری می‌یابیم که عبارت

$$\begin{aligned} S(\boldsymbol{\beta}) &= \textstyle\sum_{i=۱}^{n} \varepsilon_i^2 \\ &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

می‌نیمم شود. برای این کار کافی است از عبارت فوق نسبت به بردار $\boldsymbol{\beta}$ مشتق بگیریم یعنی

$$\frac{\partial S}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}} = -۲\mathbf{X}'\mathbf{Y} + ۲\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = ۰$$

که به صورت زیر ساده می‌شود

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y} \qquad (۳.۱)$$

معادلات (۳.۱) معادلات نرمال حداقل مربعات نامیده می‌شود. برای حل این معادلات طرفین (۳.۱) را در معکوس $\mathbf{X}'\mathbf{X}$ ضرب می‌کنیم. بنابراین برآوردگر حداقل مربعات به شرط وجود $(\mathbf{X}'\mathbf{X})^{-۱}$ به صورت زیر خواهد بود

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'\mathbf{Y} \qquad (۴.۱)$$

توجه داشته باشیم چنانچه متغیرهای توضیحی مستقل خطی باشند ماتریس $(\mathbf{X}'\mathbf{X})^{-۱}$ وجود خواهد داشت.

بردار مقادیر برازش داده شده $\hat{y}_i$ متناظر با مقادیر مشاهده شده y_i نیز به صورت زیر است

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'\mathbf{Y} \qquad (۵.۱) \\ &= \mathbf{H}\mathbf{Y}\end{aligned}$$

که در آن $\mathbf{H}$ یک ماتریس $n \times n$ است و معمولاً **ماتریس برازش** نامیده می‌شود زیرا بردار مقادیر مشاهدات را به بردار مقادیر برازش شده تصویر می‌کند.
اختلاف بین بردار مقادیر مشاهده شده و بردار مقادیر برازش داده شده بردار باقیمانده نامیده می‌شود که به صورت زیر می‌باشد

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{H}\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y}\end{aligned}$$

مجموع مربعات باقیمانده نیز به صورت زیر می‌باشد

$$\begin{aligned}SSE &= \textstyle\sum_{i=۱}^{n}(y_i - \hat{y}_i)^{۲} \\ &= \textstyle\sum_{i=۱}^{n} e_i^{۲} \\ &= \mathbf{e}'\mathbf{e}\end{aligned}$$

با جانشین کردن $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ داریم

$$\begin{aligned}SSE &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - ۲\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

چون $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$ داریم

$$SSE = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \qquad (۶.۱)$$

همان‌طور که می‌دانیم تغییرپذیری کل مشاهدات را می‌توان با مجموع مربعات تصحیح شده $S_{yy} = \sum_{i=۱}^{n}(y_i - \bar{y})^۲$ نشان داد. از طرفی می‌توان نوشت

$$y_i - \bar{y}_i = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

با استفاده از تساوی فوق به راحتی می‌توان نشان داد

$$S_{yy} = \sum_{i=۱}^{n}(\hat{y}_i - \bar{y})^۲ + \sum_{i=۱}^{n}(y_i - \hat{y}_i)^۲$$

مؤلفه اوّل در سمت راست تساوی فوق بیانگر آن بخش از تغییرپذیری است که توسط خط رگرسیون توضیح داده می‌شود و مجموع مربعات رگرسیون نامیده می‌شود. مؤلفه دوم نیز تغییرات باقیمانده که به وسیله خط رگرسیون توضیح داده نمی‌شود را اندازه‌گیری می‌کند و همان‌طور که گفته شد مجموع مربعات باقیمانده نام دارد.

بنابراین می‌توان نوشت

$$S_{yy} = SSR + SSE$$

از آنجا که

$$S_{yy} = \mathbf{Y}'\mathbf{Y} - \frac{۱}{n}(\sum_{i=۱}^{n} y_i)^۲$$

بنابراین مجموع مربعات رگرسیونی نیز به صورت زیر خواهد بود

$$SSR = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \frac{۱}{n}(\sum_{i=۱}^{n} y_i)^۲$$

خواص برآوردگر حداقل مربعات

ابتدا اریبی برآوردگر حداقل مربعات را مورد بررسی قرار می‌دهیم.

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E((\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'\mathbf{Y}) \\ &= E((\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})) \\ &= E((\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} \end{aligned}$$

بنابراین برآوردگر حداقل مربعات یک برآوردگر نااریب برای بردار پارامتر β می‌باشد. همچنین ماتریس کوواریانس $\hat{\beta}$ نیز به صورت زیر است

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= V((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

یکی از خواص مهم برآوردگر حداقل مربعات در قضیه‌ای به نام **قضیه گوس**[1]**–مارکوف**[2] ارائه می‌شود که ما آن را بدون اثبات می‌پذیریم.

قضیه ۱.۳.۱. *اگر خطاهای تصادفی ϵ دارای میانگین صفر و واریانس ثابت σ^2 بوده و دوبه‌دو ناهمبسته باشند آنگاه برآوردگر حداقل مربعات برای پارامتر β نااریب بوده و در بین تمام برآوردگرهای خطی دارای کمترین واریانس است.*

مقیاس‌سازی متغیرها

هنگامی که متغیرهای توضیحی دارای واحدهای اندازه‌گیری متفاوتی هستند معمولاً مقیاس‌سازی آنها بسیار مفید خواهد بود. برای این منظور در اینجا دو روش مقیاس‌سازی را مورد بررسی قرار می‌دهیم.

اولین روش **مقیاس‌سازی نرمال واحد** نام دارد و به صورت زیر است

$$\begin{aligned} z_{ij} &= \frac{x_{ij} - \bar{x}_j}{S_j} \qquad i = 1, 2, \ldots, n \; ; \; j = 1, 2, \ldots, p \\ y_i^* &= \frac{y_i - \bar{y}}{S_y} \qquad\qquad i = 1, 2, \ldots, n \end{aligned}$$

که در آن

$$\begin{aligned} S_j^2 &= \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1} \qquad \forall j \\ S_y^2 &= \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} \end{aligned}$$

به ترتیب واریانس‌های نمونه‌ای jامین متغیر توضیحی و متغیر پاسخ می‌باشند. در این صورت متغیرهای استاندارد شده دارای میانگین نمونه‌ای صفر و واریانس نمونه‌ای برابر واحد می‌باشند. در این صورت مدل

[1] Gauss

[2] Markov

رگرسیونی که با این متغیرهای استاندارد ساخته می‌شود بدون عرض از مبدأ خواهد بود. در این صورت برآوردگر حداقل مربعات ضرایب رگرسیونی استاندارد به صورت زیر به دست می‌آید

$$\hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}^* \qquad (۷.۱)$$

نوع دیگر مقیاس‌سازی که **مقیاس‌سازی طول واحد** نام دارد به صورت زیر می‌باشد

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{S_{jj}^{1/2}} \qquad i = 1, 2, \ldots, n \; ; \; j = 1, 2, \ldots, p$$
$$y_i^o = \frac{y_i - \bar{y}}{S_{yy}^{1/2}} \qquad i = 1, 2, \ldots, n$$

که در آن

$$S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \qquad \forall j$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

به ترتیب مجموع مربعات تصحیح شده مربوط به متغیرهای توضیحی و متغیر پاسخ می‌باشد. در این مقیاس‌سازی متغیرهای استاندارد دارای میانگین صفر و طول واحد ($\sqrt{\sum_{i=1}^{n}(w_{ij} - \bar{w}_j)^2} = 1$) می‌باشند. برآوردگر حداقل مربعات نیز به صورت زیر می‌باشد

$$\hat{\mathbf{b}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}^o \qquad (۸.۱)$$

در مقیاس‌سازی طول واحد، ماتریس $\mathbf{W}'\mathbf{W}$ به شکل ماتریس همبستگی است. یعنی

$$\mathbf{W}'\mathbf{W} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{12} & 1 & r_{23} & \cdots & r_{2p} \\ r_{13} & r_{23} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{1p} & r_{2p} & r_{3p} & \cdots & 1 \end{bmatrix}$$

که در آن

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{(S_{ii}S_{jj})^{1/2}} = \frac{S_{ij}}{(S_{ii}S_{jj})^{1/2}}$$

همبستگی ساده بین متغیرهای توضیحی X_i و X_j می‌باشد. به طور مشابه

$$\mathbf{W}'\mathbf{Y}^o = \begin{bmatrix} r_{۱y} \\ r_{۲y} \\ r_{۳y} \\ . \\ . \\ . \\ r_{py} \end{bmatrix}$$

که در آن

$$r_{jy} = \frac{\sum_{i=۱}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{(S_{jj}S_{yy})^{۱/۲}} = \frac{S_{jy}}{(S_{jj}S_{yy})^{۱/۲}}$$

همبستگی ساده بین متغیر توضیحی X_j و متغیر پاسخ Y است.

نکته ۲.۳.۱. *از آنجا که همواره داریم*

$$\mathbf{Z}'\mathbf{Z} = (n-۱)\mathbf{W}'\mathbf{W}$$

در نتیجه برآوردهای ضرایب رگرسیونی در (۱.۷) و (۱.۸) یکسان است. به این معنی که مهم نیست کدام روش مقیاس‌سازی را به کار ببریم زیرا هر دو روش برآوردهای یکسانی تولید می‌کنند.

نکته ۳.۳.۱. *ضرایب رگرسیونی* $\hat{\mathbf{b}}$ *ضرایب رگرسیونی استاندارد نامیده می‌شوند. ارتباط بین این ضرایب و ضرایب رگرسیون اولیه به صورت زیر است.*

$$\hat{\beta}_j = \hat{b}_j(\frac{S_{yy}}{S_{jj}})^{۱/۲} \qquad j = ۱, ۲, \ldots, p$$

$$\hat{\beta}_۰ = \bar{y} - \sum_{j=۱}^{p}\hat{\beta}_j\bar{x}_j$$

متراکم‌سازی داده‌ها

در بیشتر کاربردهای رگرسیونی بین متغیرهای توضیحی همبستگی وجود دارد و این امر باعث رخداد مسأله‌ای به نام هم‌خطی چندگانه می‌شود که در بیشتر موارد مشکلات فراوانی از جمله کاهش دقت برآوردگر حداقل مربعات را ایجاد می‌کند. در فصل دوم خواهیم دید که روش‌های زیادی برای حل مسأله هم‌خطی چندگانه وجود دارد که از آن جمله می‌توان به حذف بعضی متغیرهای توضیحی از مدل یا استفاده از

ترکیبات خطی خاص متغیرهای توضیحی در مدل اشاره کرد.

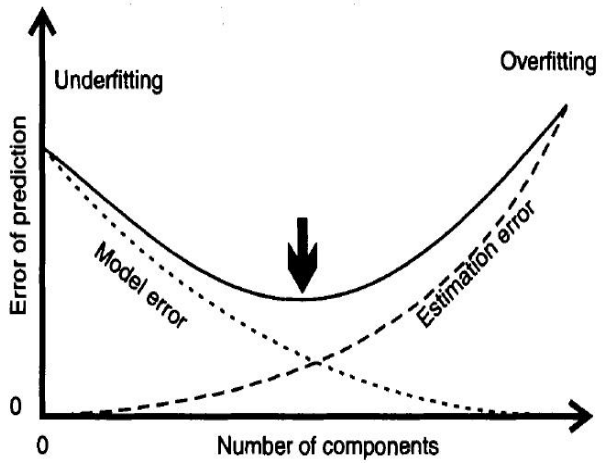
در واقع با حذف متغیرهای توضیحی از مدل و یا استفاده از ترکیبات خطی آنها در مدل نوعی متراکم‌سازی در داده‌ها ایجاد می‌کنیم. می‌توان نشان داد که متراکم‌سازی باعث ایجاد اریبی در برآوردگرهای ضرایب رگرسیونی خواهد شد. حال اگر MSE برآوردگرهای اریب حاصل از واریانس برآوردگرهای نااریب اولیه کمتر باشد می‌توان گفت: متراکم‌سازی به افزایش دقت برآوردگرهای ضرایب رگرسیون کمک کرده است.

همچنین اگر تعداد متغیرهای توضیحی در مدل زیاد باشد دقت پیشگویی کم خواهد شد. زیرا در این صورت واریانس پیشگویی $\hat{\mathbf{Y}}$ افزایش خواهد یافت. بنابراین یک مسأله مهم در متراکم‌سازی داده‌ها انتخاب **تعداد بهینه** از متغیرهایی است که باید در مدل حضور داشته باشند.

اگر تعداد متغیرهای توضیحی مورد استفاده زیاد باشد مدل حاصل به داده‌ها خیلی وابسته خواهد بود و نتایج پیشگویی ضعیفی ارائه خواهد داد. در این حالت گفته می‌شود مدل دارای **برازش اضافی** است. از طرف دیگر، چنانچه تعداد متغیرهای مورد استفاده خیلی کم باشد مدل حاصل قادر نیست بخشی از تغییرات مهم موجود در داده‌ها را توضیح دهد. در این حالت گفته می‌شود مدل دارای **کم برازشی** است.

اگر تعداد متغیرهای مورد استفاده زیاد باشد از آنجا که بیشتر تغییرات $\mathbf{X}$ مدل‌بندی می‌شود **خطای مدل‌بندی** کاهش خواهد یافت. در حالی که به علت افزایش تعداد پارامترهای مورد برآورد **خطای برآورد** افزایش خواهد داشت. این حالت در شکل (۱.۱) نشان داده شده است.
همان‌طور که مشاهده می‌شود هم خطای مدل‌بندی و هم خطای برآورد در مقدار خطای حاصل از پیشگویی مشارکت دارند. مقدار بهینه برای انتخاب تعداد متغیرهای توضیحی در بین این دو انتها یعنی حداکثر خطای برآورد و حداکثر خطای مدل قرار دارد که در شکل با پیکان نشان داده شده است.

در فصل‌های بعد بعضی از روش‌های رگرسیونی اریب را معرفی خواهیم کرد که با استفاده از متراکم سازی داده‌ها مشکل هم‌خطی را حل کرده و برآوردگرهایی ایجاد می‌کنند که دارای دقت بیشتری نسبت به برآوردگرهای حداقل مربعات هستند.



شکل ۱.۱ : انتخاب تعداد بهینه متغیرها برای جلوگیری از رخداد بیش برازش یا کم برازش

۲.۳.۱ تحلیل مؤلفه‌های اصلی

فرض کنید سیستمی شامل p متغیر تصادفی $X_1, X_2, \ldots, X_p$ می‌باشد. واضح است که برای مطالعه‌ی تغییر پذیری کل این سیستم p مؤلفه لازم است ولی اغلب می‌توان این تغییر پذیری را با تعدادی کمتر مثلاً k مؤلفه‌ی اصلی نیز بیان نمود. در این صورت میزان اطلاعی که در k مؤلفه‌ی اصلی وجود دارد تقریباً با میزان اطلاع موجود در p متغیر اولیه برابر خواهد بود. بنابراین k مؤلفه‌ی اصلی را می‌توان به جای p متغیر اولیه به کار برد و مجموعه داده‌های اولیه که شامل n اندازه روی p متغیر است را به مجموعه‌ای از داده‌های شامل n اندازه روی k مؤلفه‌ی اصلی کاهش داد. در واقع می‌توان گفت: تحلیل مؤلفه‌های اصلی که ما به اختصار آن را PCA [۳] می‌نامیم سعی در کاهش حجم داده‌ها و در نتیجه تفسیر راحت‌تری از آنها دارد.

مؤلفه‌های اصلی از نظر جبری، به صورت ترکیبات خطی ویژه‌ای از p متغیر تصادفی $X_1, X_2, \ldots, X_p$ می‌باشند. به عبارت دیگر، تحلیل مؤلفه‌های اصلی در جستجوی آن ترکیبات خطی که بیشترین واریانس را دارند می‌باشد. در واقع بزرگ بودن واریانس باعث تفکیک داده‌ها شده و در نتیجه بررسی تفاوت‌ها بین داده‌ها آسان‌تر خواهد شد.

[۳] Principal Components Analysis