





دانشگاه فردوسی مشهد
دانشکده علوم ریاضی
گروه آمار

پایان نامه

برای دریافت درجه کارشناسی ارشد در رشته
آمار اقتصادی و اجتماعی

عنوان

نمودارهای HE در مدل های خطی چندمتغیره

استاد راهنما

دکتر مجید سرمد

استاد مشاور

دکتر آرزو حبیبی راد

نگارنده

ساراشهامتی زبیدی



بسمه تعالی
مشخصات پایان‌نامه تحصیلی دانشجویان
دانشگاه فردوسی مشهد

عنوان: نمودارهای HE در مدل‌های خطی چندمتغیره

نام نویسنده: سارا شهامتی زبیدی
استاد راهنما: دکتر مجید سرمد
استاد مشاور: دکتر آرزو حبیبی‌راد

دانشکده: علوم ریاضی گروه: آمار رشته تحصیلی: آمار اقتصادی و اجتماعی

تاریخ تصویب: ۱۳۹۱/۲/۱۹ تاریخ دفاع: ۱۳۹۲/۱۱/۹

مقطع تحصیلی: کارشناسی ارشد تعداد صفحات: ۹۶

چکیده پایان‌نامه: مدل‌های خطی که شامل انواع رگرسیون، تحلیل واریانس و تحلیل کوواریانس می‌باشد، در تحقیقات کاربردی مورد استفاده قرار می‌گیرند. بسط این مدل پایه شامل تمام مدل‌های تعمیم‌یافته مانند رگرسیون چندمتغیره، رگرسیون پواسون، رگرسیون لجستیک و لگاریتم خطی است. روش‌های تشخیصی و گرافیکی نقش مهمی در بررسی مدل‌های خطی دارند. این روش‌ها کمک شایانی به پژوهشگر برای درک روابط و ویژگی متغیرها می‌کنند. با این حال روش‌های گرافیکی برای متغیرهای وابسته چندمتغیره، توسعه‌ی چندانی نیافته و یا حداقل به‌طور کامل شناخته نشده و یا استفاده نمی‌شود. نمودارهای HE که در صدد معرفی آن هستیم مربوط به کاربرد بیضی‌های داده برای نشان دادن تغییرات در فرضیه‌های صفر چندمتغیره (H)، نسبت به تغییرات خطا (E) می‌باشد.

واژگان کلیدی: مدل‌های خطی چندمتغیره، رگرسیون چندمتغیره، بیضی داده، نمودار HE

امضای استاد راهنما: تاریخ:

اظهارنامه

عنوان پایان نامه : نمودارهای HE در مدل‌های خطی چندمتغیره

اینجانب سارا شهامتی زبیدی دانشجوی دوره کارشناسی ارشد دانشکده علوم ریاضی دانشگاه فردوسی مشهد نویسنده پایان نامه تحت راهنمایی دکتر مجید سرمد متعهد می‌شوم:

- آ. تحقیقات در این رساله توسط اینجانب انجام شده و از صحت و اصالت برخوردار است.
- ب. در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.
- ج. مطالب مندرج در این پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی به جایی ارائه نشده است.
- د. کلیه حقوق این اثر متعلق به دانشگاه فردوسی مشهد است و مقالات مستخرج با نام "دانشگاه فردوسی مشهد" و یا "Ferdowsi University of Mashhad" به چاپ خواهد رسید.
- ه. حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی رساله تاثیرگذار بوده‌اند در مقالات مستخرج از آن رعایت شده است.
- و. در کلیه مراحل انجام این رساله، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده، ضوابط و اصول اخلاقی رعایت شده است.
- ز. در کلیه مراحل انجام این رساله، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده، اصل رازداری، ضوابط و اصول اخلاقی انسانی رعایت شده است.

تاریخ
امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق این اثر و محصولات آن (مقالات مستخرج، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه فردوسی مشهد است. این مطلب بایستی به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج این رساله بدون ذکر مرجع مجاز نیست.

تقدیم به

پدر و مادر عزیزم

که جز صبر و گذشت از آنها هیچ ندیدم

و بمسرم که همواره پشتیبان من بوده است.

فهرست مطالب

۶	تعاریف و مقدمات	۱
۶ مقدمه	۱.۱
۶ تعاریف جبری	۲.۱
۹ تعاریف آماری	۳.۱
۱۰ الگوی رگرسیون خطی چندگانه (کلاسیک)	۱.۳.۱
۱۲ رگرسیون چندگانه چندمتغیره	۲.۳.۱
۲۰ با داده‌های گمشده چه باید کرد؟ معرفی روش‌های جانپی	۳.۳.۱
۲۳	نمودارهای HE در مدل‌های خطی چندمتغیره	۲
۲۳ مقدمه	۱.۲
۲۴ نمودارهای دو متغیره در داده‌های چندمتغیره	۲.۲
۲۴ بیضی داده	۱.۲.۲
۲۷ دو نموداره	۲.۲.۲
۲۹ نمودارهای HE	۳.۲.۲
۴۱ نمودارهای HE برای تحلیل رگرسیون چندگانه چندمتغیره	۴.۲.۲
۴۲ نتیجه	۳.۲
۴۵	معرفی بسته توابع $heplots$	۳
۴۵ مقدمه	۱.۳
۴۶ معرفی توابع	۲.۳
۴۶ $coefplot$: رسم نمودارهای ضرایب برای مدل‌های خطی چندمتغیره	۱.۲.۳

۴۸	ellipse3d.axes : رسم محورهای بیضی‌گون سه‌بعدی	۲.۲.۳
۵۰	etasq : تعیین اندازه وابستگی جزئی (اتا-دو) برای مدل‌های خطی	۳.۲.۳
۵۳	heplot : رسم نمودارهای HE در حالت دو بعدی	۴.۲.۳
۶۰	heplot1d : رسم نمودارهای HE در حالت یک بعدی	۵.۲.۳
۶۲	heplot3d : رسم نمودارهای HE در حالت سه بعدی	۶.۲.۳
۶۳	markH0 : مشخص کردن مکان مربوط به فرضیه صفر در نمودار HE	۷.۲.۳
۶۵	pairs.mlm : رسم ماتریس نموداری HE	۸.۲.۳
۶۷	statList : بررسی آماره‌ها برای سطوح عامل‌ها	۹.۲.۳
۷۰	۴ تحلیل و تفسیر دو نمونه‌ی واقعی	
۷۰	مقدمه	۱.۴
۷۱	نمونه‌ی اول	۲.۴
۷۱	تحلیل رگرسیونی	۱.۲.۴
۷۲	نمودار HE	۲.۲.۴
۷۴	ماتریس نموداری HE	۳.۲.۴
۷۶	نمودار HE در حالت سه بعدی	۴.۲.۴
۷۷	نمونه‌ی دوم	۳.۴
۷۸	معرفی متغیرها	۱.۳.۴
۷۸	تحلیل رگرسیونی	۲.۳.۴
۷۹	نمودار HE	۳.۳.۴
۸۲	ماتریس نمودار HE:	۴.۳.۴
۸۴	نتیجه	۴.۴
۸۶	آ معرفی داده‌های موجود در بسته heplots	
۸۶	Adopted : کودکان تحت سرپرستی	۱.آ
۸۷	FootHead : اندازه‌گیری‌های دور سر در بازیکنان فوتبال	۲.آ
۸۷	Headache : آزمایشی بر روی حساسیت نسبت به صدا در افراد مبتلا به سردرد	۳.آ
۸۸	Plastic : داده‌های فیلم پلاستیکی	۴.آ
۸۹	Rohwer	۵.آ

۸۹	RootStock : اندازه‌گیری رشد درختان سیب با ریشه‌های مختلف	۶.آ
۹۰	schooldata : داده‌های مدرسه	۷.آ
۹۱	Skulls : اندازه جمجمه مصریان	۸.آ
۹۱	SocGrades : نمرات در دوره جامعه شناسی	۹.آ
۹۲	VocabGrowth : داده‌های مربوط به پیشرفت دامنه لغات	۱۰.آ
۹۳	WeightLoss : داده‌های کاهش وزن	۱۱.آ

۹۴

مراجع

فهرست شکل‌ها

- ۱.۲ بیضی داده استاندارد: نمایش انحراف معیار هر متغیر (s_1, s_2)، میانگین‌های هر
- ۲۵ . متغیر (نقطه توپر سیاه)، خط رگرسیونی Y_2 بر Y_1 و همبستگی بین متغیرها (r)
- ۲.۲ ماتریس نمودار پراکنش در داده‌های کلاسیک اندرسن: نمایشی از بیضی‌های داده با پوشانندگی %۶۸ و خطوط رگرسیون برای هر گونه از گلها، علامت مثلث برای گونه اول (*setosa*)، علامت + برای گونه دوم (*versicolor*) و علامت مربع برای گونه سوم (*virginica*)
- ۲۷
- ۳.۲ دونموداره داده‌های اندرسن، نشان‌دهنده‌ی مشاهدات (نقاط) و متغیرها (بردارها)، به همراه بیضی‌های داده %۶۸ برای گونه‌ها (*setosa*: آبی (۱)؛ *versicolor*: سبز (۲)؛ *virginica*: قرمز (۳)) و برای همه‌ی گونه‌ها (بیضی خاکستری)
- ۲۸
- ۴.۲ نمایش ایده‌های لازم برای آزمون‌های چندمتغیره با نمایش ماتریس‌های فرضیه (H) و خطا (E) برای طرح تحلیل واریانس یک طرفه با دو متغیر وابسته
- ۳۴
- ۵.۲ انواع نمودارهای HE: رابطه‌ی بین طول کاسبرگ و طول گلبرگ در داده‌های گل زنبق
- ۳۵
- ۶.۲ ماتریس نمودار HE، داده‌های کلاسیک اندرسن
- ۳۷
- ۷.۲ نمودار HE برای داده‌های سفال، نمایش آهن و آلومینیوم
- ۳۸
- ۸.۲ ماتریس نمودار HE برای داده‌های سفال. میانگین‌های گروه برای کوره‌ها توسط نام لاتین آن‌ها مشخص شده‌است.
- ۳۹
- ۹.۲ بیضی‌های داده و نمودار HE برای طرح دوطرفه
- ۴۰
- ۱.۳ نمودار ضرایب
- ۴۸
- ۲.۳ رسم محورهای بیضی‌گون در حالت سه‌بعدی
- ۵۱

۵۸	نمودار HE برای داده‌های Rohwer	۳.۳
۵۹	نمودار HE و آزمون کل مدل	۴.۳
۶۱	نمودار HE در حالت یک‌بعدی برای داده‌های Plastic	۵.۳
۶۳	نمودار HE در حالت سه‌بعدی برای داده‌های Plastic	۶.۳
۶۵	رسم مکان مربوط به فرضیه H0 در داده‌های VocabGrowth	۷.۳
۶۷	ماتریس نموداری HE برای داده‌های Rohwer	۸.۳
۷۳	نمودار HE برای داده‌های SocGrades	۱.۴
۷۵	ماتریس نموداری HE مربوط به داده‌های SocGrades	۲.۴
		نمودار HE در حالت سه بعدی. بیضی صورتی برای خطا، بیضی نیلی برای فرضیه	۳.۴
۷۷		کل، بیضی سبز برای pretest ، بیضی آبی برای gpa ، پاره‌خط سیاه برای boards	
۸۰	نمودار HE برای داده‌های مباشر	۴.۴
۸۳	ماتریس نمودار HE برای داده‌های مباشر	۵.۴

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. از تعاریف جبری برای اثبات قضایا در پایان فصل یک و از تعاریف آماری برای آشنایی با رگرسیون چندگانه و چندمتغیره و نحوه برخورد با داده‌های گمشده در فصل چهار، استفاده خواهد شد.

۲.۱ تعاریف جبری

در این بخش برخی از تعاریف مورد نیاز در جبرخطی آورده شده است. برای مطالعه بیشتر به [۱]، [۲]، [۳] و [۶] مراجعه شود.

همان‌طور که می‌دانیم، دترمینان ماتریس مربع $A_{p \times p}$ به این صورت تعریف می‌شود:

$$|A| = (-1)^{1+1} a_{11} |\hat{A}_{11}| + (-1)^{1+2} a_{12} |\hat{A}_{12}| + \dots + (-1)^{1+p} a_{1p} |\hat{A}_{1p}| = \sum_{j=1}^p (-1)^{1+j} a_{1j} |\hat{A}_{1j}|$$

که در آن \hat{A}_{ij} ماتریس حاصل از حذف سطر i ام و ستون j ام می‌باشد.

تعریف ۱.۲.۱. فرض کنید $A_{m \times n}$ و $B_{p \times q}$ دو ماتریس باشند. در این صورت حاصل ضرب کرونگر^۱ (حاصل ضرب مستقیم) A و B به صورت $A \otimes B = (a_{ij}B)$ که یک ماتریس $mp \times nq$ است، تعریف می‌شود. و در آن $A = (a_{ij})$ ، $i = 1, 2, \dots, m$ و $j = 1, 2, \dots, n$ است.

تعریف ۲.۲.۱. فرض کنید A یک ماتریس مربع $p \times p$ و I نیز ماتریس همانی از همان مرتبه باشد. اسکالرهای $\lambda_1, \lambda_2, \dots, \lambda_p$ که در معادله‌ی چندجمله‌ای $|A - \lambda I| = 0$ صدق می‌کنند را مقادیر ویژه^۲ ماتریس A می‌نامند.

تعریف ۳.۲.۱. گوییم ماتریس مربع A دارای یک مقدار ویژه‌ی λ همراه با بردار ویژه^۳ $x \neq 0$ است، هرگاه

$$Ax = \lambda x$$

شایان ذکر است که بردارهای ویژه یکتا هستند مگر این که 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.

$$(\lambda_1, x_1), (\lambda_2, x_2), \dots, (\lambda_p, x_p)$$

تعریف ۴.۲.۱. بردارهای ویژه را می‌توان طوری انتخاب کرد که بر هم عمود باشند و طول آن‌ها نیز برابر با یک باشد. به این کار استانداردسازی و بردارهای حاصل را بردارهای یکه متعامد می‌گویند. در این صورت بردارهای استاندارد شده که ما آن‌ها را با e_i نمایش می‌دهیم در شرایط زیر صدق خواهند کرد.

$$\begin{aligned} e_i' e_i &= 1 & i &= 1, 2, \dots, p \\ e_i' e_j &= 0 & i &\neq j \end{aligned}$$

نکته ۵.۲.۱. اگر ماتریس A متقارن باشد مقادیر ویژه‌ی آن حقیقی و بردارهای ویژه‌ی آن متعامد خواهند بود.

تعریف ۶.۲.۱. مجموع اعضای قطر اصلی یک ماتریس مربعی مانند A را اثر ماتریس A می‌نامند و با نماد $tr(A)$ نمایش می‌دهند.

^۱Kronocker product ^۲Eigen values ^۳Eigen vector

تعریف ۷.۲.۱. فرض کنید A یک ماتریس $p \times p$ و X یک بردار $p \times 1$ باشد. آنگاه معادلات زیر برقرار می‌باشد.

.۱

$$X'AX = tr(X'AX) = tr(AXX')$$

.۲

$$tr(A) = \sum_{i=1}^p \lambda_i$$

تعریف ۸.۲.۱. بردارهای X_1, X_2, \dots, X_p به صورت خطی وابسته هستند؛ هرگاه مجموعه‌ای از مقادیر ثابت t_1, t_2, \dots, t_p که همه صفر نیستند وجود داشته باشند به طوری که

$$\sum_{j=1}^p t_j X_j = 0$$

نکته ۹.۲.۱. اگر بردارهای X_1, X_2, \dots, X_p وابسته خطی نباشند، آنگاه مستقل خطی خواهند بود.

تعریف ۱۰.۲.۱. ماتریس A را با ستون‌های A_1, A_2, \dots, A_p در نظر بگیرید. رتبه‌ی A برابر با بیشینه تعداد بردارهای مستقل خطی در مجموعه‌ی A_1, A_2, \dots, A_p است.

تعریف ۱۱.۲.۱. ماتریس مربع A ناتکین نامیده می‌شود هرگاه رتبه‌ی آن برابر تعداد ستون‌ها (سطرها) باشد. در واقع شرط ناتکین بودن به این معنی است که ستون‌های (سطرهای) ماتریس مستقل خطی باشند.

تعریف ۱۲.۲.۱. تجزیه طیفی یک ماتریس متقارن $p \times p$ مانند A به صورت زیر می‌باشد.

$$A = \sum_{i=1}^p \lambda_i e_i e_i'$$

که در آن λ_i ها مقادیر ویژه و e_i ها بردارهای ویژه استاندارد شده هستند.

تعریف ۱۳.۲.۱. وقتی ماتریس متقارن $A_{p \times p}$ ، طوری است که برای هر $\mathbf{x}' = [x_1, x_2, \dots, x_p] \neq \mathbf{0}$

$$\mathbf{x}' A \mathbf{x} \geq 0 \quad (1.1)$$

A را معین نامنفی می‌گوییم. اگر در عبارت ۱.۱، فقط برای بردار $\mathbf{x}' = [0, 0, \dots, 0]$ تساوی برقرار شود، آن‌گاه A را معین مثبت می‌گوییم. بنابراین A معین مثبت است هرگاه برای هر بردار $\mathbf{x} \neq \mathbf{0}$

$$\mathbf{x}' A \mathbf{x} > 0$$

چون $\mathbf{x}' A \mathbf{x}$ ، فقط شامل جملات توان دوم (x_i^2) و جملات حاصل ضرب $(x_i x_k)$ است، آن‌را فرم درجه دوم می‌نامند.

نکته ۱۴.۲.۱. با بکار بردن تجزیه طیفی نسبتاً آسان، نشان داده می‌شود که یک ماتریس متقارن $A_{p \times p}$ یک ماتریس معین مثبت است اگر و فقط اگر هر مقدار ویژه‌ی ماتریس A مثبت باشد و این ماتریس معین نامنفی است اگر و فقط اگر تمام مقادیر ویژه‌اش بزرگتر یا مساوی صفر باشد.

تعریف ۱۵.۲.۱. ماتریس معین مثبت متقارن $B_{p \times p}$ و اسکالر $b > 0$ داده شده، در این صورت برای هر $(\Sigma)_{p \times p}$ معین مثبت داریم:

$$\frac{1}{|\Sigma|_b} e^{-\text{tr}((\Sigma)^{-1} B)} \leq \frac{1}{|B|_b} (2b)^{pb} e^{-bp}$$

و تساوی فقط وقتی برقرار است که:

$$\Sigma = \frac{1}{2b} B.$$

۳.۱ تعاریف آماری

در این بخش برای آشنایی بیشتر با رگرسیون چندگانه و چندمتغیره و بررسی تفاوت این دو مدل به معرفی آن‌ها پرداخته می‌شود. مطالب این بخش از [۲]، [۵]، [۸] و [۱۳] برداشت شده است.

۱.۳.۱ الگوی رگرسیون خطی چندگانه (کلاسیک)

فرض کنید p, x_1, x_2, \dots, x_p متغیر مستقل، در ارتباط با یک متغیر وابسته y باشد. برای مثال با

قیمت فصلی منزل در بازار $y =$

مساحت محل زندگی به متر مربع $x_1 =$

محل (نشان‌دهنده‌ی منطقه شهری) $x_2 =$

قیمت ارزیابی شده سال گذشته $x_3 =$

کیفیت ساختمان (قیمت به متر مربع) $x_4 =$

الگوی رگرسیون خطی کلاسیک، بیان می‌کند که y از یک میانگین که به صورت خطی به x_i ها و خطای تصادفی ε که به خاطر خطای اندازه‌گیری و اثرات سایر متغیرهایی که در الگو منظور نشده‌اند، وابسته است ساخته می‌شود. مقادیر متغیرهای مستقل که توسط محقق از آزمایش به دست می‌آید، ثابت تلقی می‌شوند. خطا (و در نتیجه متغیر وابسته) به صورت یک متغیر تصادفی در نظر گرفته می‌شود که رفتار آن به وسیله مجموعه‌ای از فرض‌های توزیع خاص مشخص می‌شود. به ویژه یک الگوی رگرسیون خطی با یک متغیر وابسته به این صورت است:

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

این الگو به عنوان رگرسیون چندگانه معرفی می‌شود. اصطلاح خطی به این حقیقت برمی‌گردد که میانگین Y ، تابعی خطی از پارامترهای نامعلوم $\beta_0, \beta_1, \dots, \beta_p$ است. با داشتن n مشاهده مستقل روی y و مقادیر مربوط x_i ، الگوی کامل به صورت زیر در می‌آید:

$$Y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + \varepsilon_2$$

⋮

$$Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + \varepsilon_n \quad (2.1)$$

که در آن جملات خطا باید دارای خواص زیر باشند:

.۱

$$E(\varepsilon_j) = 0$$

.۲

$$\text{Var}(\varepsilon_j) = \sigma^2 \text{ (ثابت)}$$

.۳

$$\text{Cov}(\varepsilon_j, \varepsilon_k) = 0 \quad j \neq k \quad j, k = 1, \dots, n \quad (3.1)$$

(پارامترهای β و σ^2 نامعلوم هستند.)

و با نماد ماتریسی به صورت زیر نوشته می‌شود:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

یا

$$\mathbf{Y}_{(n \times 1)} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\varepsilon}_{(n \times 1)}$$

و خواصی که در ۳.۱ بیان شد. به صورت زیر درمی‌آید:

.۱

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}_{(n \times 1)}$$

.۲

$$\text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}_{n \times n}$$

و پارامترهای β و σ^2 نامعلوم هستند.

توجه داشته باشید که وجود یک در ستون اول ماتریس طرح \mathbf{X} برای جمله ثابت β_0 است. کافی است متغیر تصنعی $x_{j_0} = 1$ را معرفی کنیم. لذا

$$\beta_0 + \beta_1 x_{j_1} + \beta_2 x_{j_2} + \dots + \beta_p x_{j_p} = \beta_0 x_{j_0} + \beta_1 x_{j_1} + \dots + \beta_p x_{j_p}$$

هر ستون در ماتریس \mathbf{X} شامل n مشاهده مربوط به مقدار متغیر مستقل مربوطه است. در حالی که سطر j ام \mathbf{X} مقادیر مربوط به تمام متغیرهای مستقل در آزمایش j ام را شامل می‌شود. اگرچه فرض‌های جمله خطا در ۳.۱ بسیار لازم است ولی در مباحث فواصل اطمینان و آزمون فرضیه‌ها فرض نرمال بودن توأم هم به فرضیات افزوده می‌شود.

۲.۳.۱ رگرسیون چندگانه چندمتغیره

در این قسمت الگو و مفهوم رگرسیون چندمتغیره بیان خواهد شد. در بخش ۱.۳.۱ تعریفی برای رگرسیون چندگانه معرفی شد. و این همان چیزی است که برخی از کاربران آمار که آشنایی کمی با آمار دارند آن را به اشتباه به عنوان الگوی رگرسیون چندمتغیره می‌شناسند. ولی در بحث‌های تخصصی رگرسیون، زمانی از رگرسیون چندمتغیره صحبت می‌شود که بیش از یک متغیر وابسته وجود داشته باشد. به عبارت دیگر بررسی رابطه‌ی بین یک یا چند متغیر مستقل با چند متغیر وابسته مدنظر است. الگوی رگرسیون چندمتغیره در حالت کلی ممکن است بسیار پیچیده باشد. در اینجا فقط الگوی رگرسیونی چندمتغیره خطی معرفی خواهد شد. یعنی رابطه‌ی بین متغیر(های) مستقل با متغیرهای وابسته، خطی فرض می‌شود. بنابراین واضح است که رگرسیون خطی چندگانه حالت خاصی از رگرسیون خطی چندمتغیره است. در این قسمت عمدتاً از مطالب [۲]، [۸] و [۱۴] استفاده شده است.

رگرسیون خطی چندمتغیره

رگرسیون چندگانه به راحتی به مدلی گسترش پیدا می‌کند که شامل $m > 1$ متغیر وابسته است. در این قسمت، مدل‌بندی رابطه‌ی بین m متغیر وابسته و p متغیر مستقل را که هر کدام از m متغیر وابسته از رابطه‌ی خاص خودش پیروی می‌کند، بررسی می‌کنیم.

موارد زیر به عنوان مثال‌هایی از رگرسیون چندمتغیره قلمداد می‌شوند.

۱. مقیاس‌های اندازه‌گیری برای $m = 5$ آلاینده هوا در $n = 42$ روز مختلف، در اینجا ۵ متغیر

وابسته و برای هر آلایند $p = ۲$ پارامتر قدرت باد و شدت نور خورشید را به عنوان متغیر مستقل داریم.

۲. نمرات امتحانی برای یک کلاس درس؛ نمرات در $m = ۳$ درس مختلف به عنوان متغیرهای وابسته، و $p = ۳$ پارامتر سن، جنس و درآمد برای هر دانش‌آموز به عنوان متغیرهای مستقل، در یک کلاس $n = ۳۰$ نفره.

فرض می‌کنیم هر متغیر وابسته یک رابطه خطی با همه متغیرهای مستقل داشته باشد به طوری که رابطه‌ی توأم متغیرهای وابسته Y_1, Y_2, \dots, Y_m با متغیرهای مستقل x_1, x_2, \dots, x_p به وسیله رابطه‌های زیر توجیه شود:

$$\begin{cases} Y_1 = \beta_{01} + \beta_{11}x_1 + \dots + \beta_{p1}x_p + \varepsilon_1 \\ Y_2 = \beta_{02} + \beta_{12}x_1 + \dots + \beta_{p2}x_p + \varepsilon_2 \\ \dots \\ Y_m = \beta_{0m} + \beta_{1m}x_1 + \dots + \beta_{pm}x_p + \varepsilon_m \end{cases}$$

از آن‌جا که متغیرهای وابسته Y_1, Y_2, \dots, Y_m خودشان معمولاً از هم مستقل نیستند. یک همبستگی یا ساختار کوواریانس بین آن‌ها برقرار است. جمله‌ی برداری خطای $\varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m]'$ دارای میانگین و واریانس به ترتیب $E(\varepsilon) = 0$ و $Var(\varepsilon) = \Sigma$ است. از این رو جملات خطای مرتبط با وابسته‌های مختلف، ممکن است همبسته باشند. بدیهی است که این بردار با بردار ε در رگرسیون چندگانه تفاوت ساختاری دارد. برای پایه‌گذاری نمادهایی که با الگوی رگرسیون خطی کلاسیک مطابقت کند، برای یک نمونه تصادفی به حجم n فرض کنید $[x_{j0}, x_{j1}, \dots, x_{jp}]$ مقادیر متغیرهای مستقل برای آزمایش یا فرد j ام را نشان دهد. فرض می‌کنیم $\mathbf{Y}_j = [Y_{j1}, Y_{j2}, \dots, Y_{jm}]$ مقادیر متغیرهای وابسته و $\varepsilon_j = [\varepsilon_{j1}, \varepsilon_{j2}, \dots, \varepsilon_{jm}]$ بردار خطاها برای فرد j ام باشند، $(j = 1, 2, \dots, n)$. ماتریس طرح که شامل مقادیر متغیر مستقل برای افراد مختلف است به صورت زیر می‌باشد:

$$\mathbf{X}_{n \times (p+1)} = \begin{bmatrix} x_{10} & x_{11} & \dots & x_{1p} \\ x_{20} & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \dots & x_{np} \end{bmatrix}$$

با توجه به نماد ماتریسی، این ماتریس مانند ماتریس طرح مربوط به الگوی رگرسیون یک متغیره چندگانه است. توجه داریم که $x_{j_0} = 1$ است. ماتریس‌های مربوط به متغیرهای وابسته، پارامترها و جملات خطا نیز به این صورت تعریف می‌شوند:

ماتریس داده‌های متغیرهای وابسته:

$$\mathbf{Y}_{n \times m} = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1m} \\ Y_{21} & Y_{22} & \dots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nm} \end{bmatrix} = [Y_{(1)} | Y_{(2)} | \dots | Y_{(m)}]$$

در این ماتریس i امین سطر مربوط به i امین مشاهده چند متغیره، و j امین ستون مربوط به j امین متغیر اندازه گرفته شده است. و $Y_{(j)}$ معرف j امین ستون است.

ماتریس ضرایب رگرسیون:

$$\boldsymbol{\beta}_{(p+1) \times m} = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pm} \end{bmatrix} = [\beta_{(1)} | \beta_{(2)} | \dots | \beta_{(m)}]$$

ماتریس خطاهای رگرسیونی:

$$\boldsymbol{\varepsilon}_{n \times m} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & \varepsilon_{nm} \end{bmatrix} = [\varepsilon_{(1)} | \varepsilon_{(2)} | \dots | \varepsilon_{(m)}] = \begin{bmatrix} \varepsilon'_1 \\ \varepsilon'_2 \\ \vdots \\ \varepsilon'_n \end{bmatrix}$$

توجه کنید که سطرهای ماتریس $\boldsymbol{\varepsilon}$ یعنی $\varepsilon'_1, \dots, \varepsilon'_n$ جملات خطای مربوط به اعضای مختلف نمونه را نشان می‌دهد و از هم مستقل هستند ولی ستون‌های ماتریس $\boldsymbol{\varepsilon}$ یعنی $\varepsilon_{(1)}, \dots, \varepsilon_{(m)}$ متناظر m اندازه‌گیری مختلف از یک فرد یا یک واحد آماری هستند، لذا همبسته می‌باشند. در حالت خاص اگر $m = 1$ باشد، ماتریس $\boldsymbol{\varepsilon}$ همان بردار $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ در الگوی رگرسیون خطی چندگانه خواهد بود.

در نتیجه الگوی رگرسیون خطی چندمتغیره با m متغیر وابسته و p متغیر مستقل را با نماد ماتریسی می‌توان به صورت زیر نوشت:

$$\mathbf{Y}_{n \times m} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times m} + \boldsymbol{\varepsilon}_{n \times m} \quad (۴.۱)$$

که در آن

$$E(\varepsilon_{(i)}) = 0, \quad Cov(\varepsilon_{(i)}, \varepsilon_{(k)}) = \sigma_{ik} I; \quad i, k = 1, 2, \dots, m \quad (۵.۱)$$

m مشاهده، روی آزمایش j ام دارای ماتریس کوواریانس $\Sigma = \{\sigma_{ik}\}$ است ولی مشاهدات از آزمایش‌های مختلف ناهمبسته‌اند. در این جا β و σ_{ik} پارامترهای نامعلوم‌اند و هم‌چنین فرض می‌کنیم خطاهای مدل (ε) متغیرهای توأمآ نرمال با میانگین صفر هستند. از آن جا که سطرهای ماتریس ε ناهمبسته هستند ماتریس واریانس-کوواریانس یک ستون از این ماتریس، یک ماتریس قطری به صورت زیر می‌باشد که σ_{jj} ، واریانس متغیر وابسته j ام می‌باشد.

$$Cov(\varepsilon_{(j)}, \varepsilon_{(j)}) = \sigma_{jj} I = \begin{bmatrix} \sigma_{jj} & 0 & \dots & 0 \\ 0 & \sigma_{jj} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{jj} \end{bmatrix}, \forall i, j = 1, 2, \dots, m$$

همان‌طور که مشاهده می‌کنیم با توجه به نرمال بودن توزیع خطاها، این ماتریس بیان‌کننده‌ی این مطلب است که خطاهای مربوط به متغیر وابسته j ام از یکدیگر مستقل بوده و دارای واریانس ثابت σ_{jj} می‌باشند. چرا که خطاهای مربوط به متغیر وابسته j ام مربوط به اعضای مختلف نمونه است.

به عنوان مثال، در بررسی ارتباط بین وزن و قد نوزاد به عنوان متغیرهای وابسته و سن مادر به عنوان متغیر مستقل، فرض کنید مثلاً $\varepsilon_{(1)}$ بردار خطای مربوط به متغیر وابسته وزن و $\varepsilon_{(2)}$ بردار خطای مربوط به متغیر وابسته قد باشد. در این صورت ماتریس واریانس کوواریانس ۵.۱ را در نظر می‌گیریم، اگر $i = k = 1$ آنگاه ماتریس واریانس کوواریانس بیان‌کننده این مطلب است که خطاهای مربوط به متغیر وابسته وزن از یکدیگر مستقل بوده و دارای واریانس σ_{11} می‌باشند، یعنی مشاهدات