دانشکده علوم ریاضی
گروه آمار

پایان‌نامه

برای دریافت درجه کارشناسی ارشد در رشته
آمار اقتصادی و اجتماعی

عنوان

نمودارهای HE در مدل‌های خطی چندمتغیره

استاد راهنما

دکتر مجید سرمد

استاد مشاور

دکتر آرزو حبیبی‌راد

نگارنده

سارا شهامتی زیدی

بسمه تعالی

مشخصات پایان‌نامه تحصیلی دانشجویان

دانشگاه فردوسی مشهد

عنوان: نمودارهای **HE** در مدل‌های خطی چندمتغیره

نام نویسنده: سارا شهامتی زیبدی

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 پوآسون، 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 این‌حال روش‌های گرافیکی برای متغیرهای وابسته‌ی 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، بیضی‌داده، نمودارHE

امضای استاد راهنما: تاریخ:

اظهارنامه

عنوان پایان‌نامه : نمودارهای **HE** در مدل‌های خطی چندمتغیره

اینجانب سارا شهامتی زیبدی دانشجوی دوره کارشناسی ارشد دانشکده علوم ریاضی دانشگاه فردوسی مشهد نویسنده پایان‌نامه 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 پایان‌نامه 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. درکلیه مراحل انجام این رساله، در مواردی که از موجود زنده(یا بافت‌های آن‌ها) استفاده شده، ضوابط و اصول اخلاقی رعایت شده است.

ز. درکلیه 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 همسرم که همواره پشتیبان من بوده است.

فهرست مطالب

فهرست شکل‌ها

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 $\mathbf{A}_{p\times p}$به این صورت تعریف می‌شود:

$$|\mathbf{A}| = (-۱)^{۱+۱}a_{۱۱}|\hat{\mathbf{A}}_{۱۱}|+(-۱)^{۱+۲}a_{۱۲}|\hat{\mathbf{A}}_{۱۲}|+...+(-۱)^{۱+p}a_{۱p}|\hat{\mathbf{A}}_{۱p}| = \sum_{j=۱}^{p}(-۱)^{۱+j}a_{۱j}|\hat{\mathbf{A}}_{۱j}|$$

که در آن $\hat{\mathbf{A}}_{ij}$ ماتریس حاصل از حذف سطر iام و ستون jام می‌باشد.

تعریف ۱.۲.۱. *فرض کنید $\boldsymbol{A}_{m\times n}$ و $\boldsymbol{B}_{p\times q}$ دو ماتریس باشند. در این‌صورت حاصل‌ضرب کرونکر[۱] (حاصل‌ضرب مستقیم) $\boldsymbol{A}$ و $\boldsymbol{B}$ به صورت $\boldsymbol{A}\otimes\boldsymbol{B}=(a_{ij}\boldsymbol{B})$ که یک ماتریس $mp\times nq$ است، تعریف می‌شود. و در آن $\boldsymbol{A}=(a_{ij})$، $i=۱,۲,\ldots,m$ و $j=۱,۲,\ldots,n$ است.*

تعریف ۲.۲.۱. *فرض کنید $\boldsymbol{A}$ یک ماتریس مربع $p\times p$ و $\boldsymbol{I}$ نیز ماتریس همانی از همان مرتبه باشد. اسکالرهای $\lambda_۱,\lambda_۲,\ldots,\lambda_p$ که در معادله‌ی چندجمله‌ای $|\boldsymbol{A}-\lambda\boldsymbol{I}|=۰$ صدق می‌کنند را مقادیر ویژه[۲] ماتریس $\boldsymbol{A}$ می‌نامند.*

تعریف ۳.۲.۱. *گوییم ماتریس مربع $\boldsymbol{A}$ دارای یک مقدار ویژه‌ی λ همراه با بردار ویژه[۳] $\boldsymbol{x}\neq ۰$ است، هرگاه*

$$\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}$$

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.

$$(\lambda_۱,\boldsymbol{x}_۱),(\lambda_۲,\boldsymbol{x}_۲),\ldots,(\lambda_p,\boldsymbol{x}_p)$$

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 $\boldsymbol{e}_i$ نمایش می‌دهیم در شرایط زیر صدق خواهند کرد.*

$$\begin{aligned}\boldsymbol{e}_i'\boldsymbol{e}_i&=۱ \qquad i=۱,۲,\ldots,p\\ \boldsymbol{e}_i'\boldsymbol{e}_j&=۰ \qquad i\neq j\end{aligned}$$

نکته ۵.۲.۱. *اگر ماتریس $\boldsymbol{A}$ 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 $\boldsymbol{A}$ را اثر ماتریس $\boldsymbol{A}$ می‌نامند و با نماد $\boldsymbol{tr(A)}$ نمایش می‌دهند.*

[۱] Kronocker product [۲] Eigen values [۳] Eigen vector

تعریف ۷.۲.۱. *فرض کنید $\boldsymbol{A}$ یک ماتریس $p \times p$ و $\boldsymbol{x}$ یک بردار $p \times ۱$ باشد. آنگاه معادلات زیر برقرار می‌باشد.*

۱.

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = tr(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}')$$

۲.

$$tr(\boldsymbol{A}) = \sum_{i=۱}^{p} \lambda_i$$

تعریف ۸.۲.۱. *بردارهای $\boldsymbol{x}_۱, \boldsymbol{x}_۲, \ldots, \boldsymbol{x}_p$ به صورت خطی وابسته هستند؛ هرگاه مجموعه‌ای از مقادیر ثابت $t_۱, t_۲, \ldots, t_p$ که همه صفر نیستند وجود داشته باشند به‌طوری که*

$$\sum_{j=۱}^{p} t_j \boldsymbol{x}_j = ۰$$

نکته ۹.۲.۱. *اگر بردارهای $\boldsymbol{x}_۱, \boldsymbol{x}_۲, \ldots, \boldsymbol{x}_p$ وابسته خطی نباشند، آنگاه مستقل خطی خواهند بود.*

تعریف ۱۰.۲.۱. *ماتریس $\boldsymbol{A}$ را با ستون‌های $A_۱, A_۲, \ldots, A_p$ در نظر بگیرید. رتبه‌ی $\boldsymbol{A}$ برابر با بیشینه تعداد بردارهای مستقل خطی در مجموعه‌ی $A_۱, A_۲, \ldots, A_p$ است.*

تعریف ۱۱.۲.۱. *ماتریس مربع $\boldsymbol{A}$ 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 $\boldsymbol{A}$ به صورت زیر می‌باشد.*

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

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

تعریف ۱.۲.۱۳. *وقتی ماتریس متقارن $\boldsymbol{A}_{p\times p}$، طوری است که برای هر $\boldsymbol{x}' = [x_1, x_2, \ldots, x_p] \neq 0$*

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0 \qquad (1.1)$$

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

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

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

نکته ۱.۲.۱۴. *با بکار بردن تجزیه طیفی نسبتاً آسان، نشان داده می‌شود که یک ماتریس متقارن $\boldsymbol{A}_{p\times p}$یک ماتریس معین مثبت است اگر و فقط اگر هر مقدار ویژه‌ی ماتریس $\boldsymbol{A}$ مثبت باشد و این ماتریس معین نامنفی است اگر و فقط اگر تمام مقادیر ویژه‌اش بزرگتر یا مساوی صفر باشد.*

تعریف ۱.۲.۱۵. *ماتریس معین مثبت متقارن $\boldsymbol{B}_{p\times p}$ و اسکالر $b > 0$ داده شده، در این صورت برای هر $(\Sigma)_{p\times p}$ معین مثبت داریم:*

$$\frac{1}{|\sum|^b} e^{\frac{-tr((\sum)^{-1}\boldsymbol{B})}{2}} \leq \frac{1}{|\boldsymbol{B}|^b}(2b)^{pb}e^{-bp}$$

و تساوی فقط وقتی برقرار است که:

$$\sum = \frac{1}{2b}\boldsymbol{B}.$$

۳.۱ تعاریف آماری

در این بخش برای آشنایی بیشتر با رگرسیون چندگانه و چندمتغیره وبررسی 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 $x_1, x_2, \ldots, x_p$، p متغیر مستقل، در ارتباط با یک متغیر وابسته y باشد. برای مثال با

$$\begin{aligned} y &= \text{قیمت فصلی منزل در بازار} \\ x_1 &= \text{مساحت محل زندگی به متر مربع} \\ x_2 &= \text{محل (نشان‌دهنده‌ی منطقه شهری)} \\ x_3 &= \text{قیمت ارزیابی شده سال گذشته} \\ x_4 &= \text{کیفیت ساختمان (قیمت به متر مربع)} \end{aligned}$$

الگوی رگرسیون خطی کلاسیک، بیان می‌کند که y از یک میانگین که به صورت خطی به x_iها 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 به‌دست 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:

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

این الگو به عنوان رگرسیون چندگانه معرفی می‌شود. اصطلاح خطی به این حقیقت برمی‌گردد که میانگین Y، تابعی خطی از پارامترهای نامعلوم $\beta_0, \beta_1, \ldots, \beta_p$ است.

با داشتن n مشاهده مستقل روی y و مقادیر مربوط x_i، الگوی کامل به صورت زیر در می‌آید:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n \end{aligned} \qquad (۲.۱)$$

که در آن جملات خطا باید دارای خواص زیر باشند:

۱.

$$E(\varepsilon_j) = ۰$$

۲.

$$Var(\varepsilon_j) = \sigma^۲ \text{(ثابت)}$$

۳.

$$Cov(\varepsilon_j, \varepsilon_k) = ۰ \quad j \neq k \quad j,k = ۱,\ldots,n \qquad (۳.۱)$$

(پارامترهای β و $\sigma^۲$ نامعلوم هستند.)

و با نماد ماتریسی به صورت زیر نوشته می‌شود:

$$\begin{bmatrix} Y_۱ \\ Y_۲ \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} ۱ & x_{۱۱} & x_{۱۲} & \ldots & x_{۱p} \\ ۱ & x_{۲۱} & x_{۲۲} & \ldots & x_{۲p} \\ \vdots & \vdots & \vdots & & \vdots \\ ۱ & x_{n۱} & x_{n۲} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_۰ \\ \beta_۱ \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_۱ \\ \varepsilon_۲ \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

یا

$$\mathbf{Y}_{(n\times ۱)} = \mathbf{X}_{n\times(p+۱)}\boldsymbol{\beta}_{(p+۱)\times ۱} + \boldsymbol{\varepsilon}_{(n\times ۱)}$$

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

۱.

$$E(\boldsymbol{\varepsilon}) = ۰_{(n\times ۱)}$$

۲.

$$Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\sigma}^۲\mathbf{I}_{n\times n}$$

و پارامترهای β و $\sigma^۲$ نامعلوم هستند.

توجه داشته باشید که وجود یک در ستون اول ماتریس طرح $\mathbf{X}$ برای جمله ثابت $\beta_۰$ است. کافی‌است متغیر تصنعی $x_{j۰} = ۱$ را معرفی کنیم. لذا

$$\beta_۰ + \beta_۱ x_{j۱} + \beta_۲ x_{j۲} + \ldots + \beta_p x_{jp} = \beta_۰ x_{j۰} + \beta_۱ x_{j۱} + \ldots + \beta_p x_{jp}$$

هر ستون در ماتریس $\mathbf{X}$ شامل n مشاهده مربوط به مقدار متغیر مستقل مربوطه است. در حالی که سطر jام $\mathbf{X}$ مقادیر مربوط به تمام متغیرهای مستقل در آزمایش jام 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 $m > ۱$ 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 $m = ۵$ آلاینده هوا در $n = ۴۲$ روز مختلف، در اینجا ۵ متغیر

وابسته و برای هر آلاینده $p = ۲$ 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 $Y_۱, Y_۲, \ldots, Y_m$ با متغیرهای مستقل $x_۱, x_۲, \ldots, x_p$ به وسیله رابطه‌های زیر توجیه شود:

$$\begin{cases} Y_۱ = \beta_{۰۱} + \beta_{۱۱}x_۱ + \ldots + \beta_{p۱}x_p + \varepsilon_۱ \\ Y_۲ = \beta_{۰۲} + \beta_{۱۲}x_۱ + \ldots + \beta_{p۲}x_p + \varepsilon_۲ \\ \ldots \\ Y_m = \beta_{۰m} + \beta_{۱m}x_۱ + \ldots + \beta_{pm}x_p + \varepsilon_m \end{cases}$$

از آن‌جا که متغیرهای وابسته $Y_۱, Y_۲, \ldots, Y_m$ خودشان معمولاً از هم مستقل نیستند. یک همبستگی یا ساختار کوواریانس بین آن‌ها برقرار است. جمله‌ی برداری خطای $\boldsymbol{\varepsilon} = [\varepsilon_۱, \varepsilon_۲, \ldots, \varepsilon_m]'$ دارای میانگین و واریانس به ترتیب $E(\boldsymbol{\varepsilon}) = ۰$ و $Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$ است. از این رو جملات خطای مرتبط با وابسته‌های مختلف، ممکن است همبسته باشند. بدیهی است که این بردار با بردار $\boldsymbol{\varepsilon}$ 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 $[x_{j۰}, x_{j۱}, \ldots, x_{jp}]$ مقادیر متغیرهای مستقل برای آزمایش یا فرد jام را نشان دهد. فرض می‌کنیم $\mathbf{Y}_j = [Y_{j۱}, Y_{j۲}, \ldots, Y_{jm}]$ مقادیر متغیرهای وابسته و $\boldsymbol{\varepsilon}_j = [\varepsilon_{j۱}, \varepsilon_{j۲}, \ldots, \varepsilon_{jm}]$ بردار خطاها برای فرد jام باشند، $(j = ۱, ۲, \ldots, n)$. ماتریس طرح که شامل مقادیر متغیر مستقل برای افراد مختلف است به صورت زیر می‌باشد:

$$\mathbf{X}_{n\times(p+۱)} = \begin{bmatrix} x_{۱۰} & x_{۱۱} & \ldots & x_{۱p} \\ x_{۲۰} & x_{۲۱} & \ldots & x_{۲p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n۰} & x_{n۱} & \ldots & x_{np} \end{bmatrix}$$

با توجه به نماد ماتریسی، این ماتریس مانند ماتریس طرح مربوط به الگوی رگرسیون یک متغیره چندگانه است. توجه داریم که $x_{j0} = 1$ است. ماتریس‌های مربوط به متغیرهای وابسته، پارامترها و جملات خطا نیز به این صورت تعریف می‌شوند:

ماتریس داده‌های متغیرهای وابسته:

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

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

ماتریس ضرایب رگرسیون:

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

ماتریس خطاهای رگرسیونی:

$$\boldsymbol{\varepsilon}_{n\times m} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & \varepsilon_{nm} \end{bmatrix} = \left[\varepsilon_{(1)}|\varepsilon_{(2)}|\dots|\varepsilon_{(m)}\right] = \begin{bmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_n' \end{bmatrix}$$

توجه کنید که سطرهای ماتریس ε یعنی $\varepsilon_n', \dots, \varepsilon_2', \varepsilon_1'$ جملات خطای مربوط به اعضای مختلف نمونه را نشان می‌دهد و از هم مستقل هستند ولی ستون‌های ماتریس ε یعنی $\varepsilon_{(m)}, \dots, \varepsilon_{(2)}, \varepsilon_{(1)}$ متناظر m اندازه‌گیری مختلف از یک فرد یا یک واحد آماری هستند، لذا همبسته می‌باشند. در حالت خاص اگر $m = 1$، باشد، ماتریس ε 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:

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

که در آن

$$E(\varepsilon_{(i)}) = ۰, \quad Cov(\varepsilon_{(i)}, \varepsilon_{(k)}) = \sigma_{ik}I; \quad i,k = ۱, ۲, \ldots, m \qquad (۵.۱)$$

m مشاهده، روی آزمایش jام دارای ماتریس کوواریانس $\Sigma = \{\sigma_{ik}\}$ است ولی مشاهدات از آزمایش‌های مختلف ناهمبسته‌اند. در این‌جا β و σ_{ik} پارامترهای نامعلوم‌اند و هم‌چنین فرض می‌کنیم خطاهای مدل (ε) متغیرهای توأماً نرمال با میانگین صفر هستند.

از آن‌جا 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 jام می‌باشد.

$$Cov(\varepsilon_{(j)}, \varepsilon_{(j)}) = \sigma_{jj}I = \begin{bmatrix} \sigma_{jj} & ۰ & \ldots & ۰ \\ ۰ & \sigma_{jj} & \ldots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ ۰ & ۰ & \ldots & \sigma_{jj} \end{bmatrix}, \forall i,j = ۱, ۲, \ldots, m$$

همان‌طور که مشاهده می‌کنیم با توجه به نرمال بودن توزیع خطاها، این ماتریس بیان‌کننده‌ی این مطلب است که خطاهای مربوط به متغیر وابسته jام از یکدیگر مستقل بوده و دارای واریانس ثابت σ_{jj} می‌باشند. چرا که خطاهای مربوط به متغیر وابسته jام 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 $\varepsilon_{(۱)}$ بردار خطای مربوط به متغیر وابسته وزن و $\varepsilon_{(۲)}$ بردار خطای مربوط به متغیر وابسته قد باشد. در این صورت ماتریس واریانس کوواریانس ۵.۱ را در نظر می‌گیریم، اگر $i = k = ۱$ آنگاه ماتریس واریانس کوواریانس بیان کننده این مطلب است که خطاهای مربوط به متغیر وابسته وزن از یکدیگر مستقل بوده و دارای واریانس $\sigma_{۱۱}$ می‌باشند، یعنی مشاهدات