برآورد نااریب ضرایب در مدل‌های خطای اندازه‌گیری به روش کمترین مربعات تعمیم‌یافته

حبیبه خاندوزی، اسماعیل شیرازی
گروه آمار، دانشکده علوم، دانشگاه آزاد اسلامی واحد مشهد
گروه آمار، دانشکده علوم پایه، دانشگاه گنبد کاووس

چکیده

مدل‌های خطای اندازه‌گیری زمانی ارائه می‌شوند که متغیرهای به کار گرفته شده در مدل با خطا اندازه‌گیری شده باشند. در این حالت برای برآورد ضرایب مدل، روش کمترین مربعات با استفاده از مشاهدات اصلی، برآوردهایی اریب را نتیجه می‌دهد. بر این اساس در این مقاله به روش کمترین مربعات تعمیم‌یافته، برآوردگرهای نااریب معرفی می‌گردد.

واژه‌های کلیدی: روش کمترین مربعات تعمیم یافته، مدل خطای اندازه‌گیری، مدل وراساختاری.

۱ مقدمه

در اکثر آزمایش‌هایی که در سراسر دنیا صورت می‌گیرد، نتایج مشاهدات علمی، مقادیر واقعی نبوده بلکه مقادیر بدست آمده همراه با یک نوع خطا است. در علوم آماری چنین خطایی به خطای اندازه‌گیری معروف است. به خصوص در علومی مانند علوم آموزشی، علوم زیستی، خطای اندازه‌گیری غیر قابل انکار است به عنوان مثال در تعیین سطح نمرات یک کلاس یا تعیین ضریب هوشی افراد، بسته به موقعیت زمانی، نوع سوالات و وضعیت روحی افراد، نمی‌توان به صورت دقیق مقادیر واقعی و دقیق آنها را اندازه‌گیری کرد.

این نوع خطا ممکن است به دلایل مختلفی از جمله دقت پایین دستگاه اندازه‌گیری، شرایط نامناسب زمان آزمایش و اشتباه در ثبت اطلاعات رخ دهد. چنین خطایی باعث بروز مشکلاتی در تحلیل داده‌ها شده و می‌تواند اعتبار نتایج بدست آمده را زیر سوال ببرد و حتی نادیده گرفتن چنین خطایی معمولاً باعث نتایج گمراه کننده‌ای

میشود. از نقطه نظر تاریخی هنگامی که فولر[۱] [۶] قصد داشت مدل رگرسیونی نیتروژن موجود در خاک را بررسی کند، متوجه این نوع خطا شد. وی بعد از آزمایش‌های دقیق متوجه شد که نیتروژن موجود در خاک را نتوانسته به صورت دقیق اندازه‌گیری کند. سپس نشان داد که با نادیده گرفتن خطای اندازه‌گیری، پارامترهای مدل رگرسیونی اریب برآورد شده و در نتیجه باعث برازش مدل‌های گمراه کننده‌ای شده است. بعلاوه ایشان مطالعات گسترده‌ای روی مدل‌های رگرسیونی ساده که متغیرهای رگرسیونی یا متغیر پاسخ یا هر دوی آنها آلوده به خطای اندازه‌گیری است انجام داد. حاصل تلاش‌های وی منجر به نگارش کتاب جامعی به نام مدل‌های خطای اندازه‌گیری[۲] شد. مدل خطای اندازه‌گیری که از آن به اختصار با عنوان مدل ME یاد می‌کنیم،به عنوان یک تعمیم از مدل رگرسیون استاندارد به شمار می‌آید و زمانی مورد بحث قرار می‌گیرد که هر دو متغیر پاسخ و متغیر رگرسیونی تحت بررسی توأم با خطا اندازه‌گیری شده باشند. در این حالت نیز هدف برازش یک خط رگرسیونی مستقیم بین دو متغیر پاسخ (y) و رگرسیونی (x_i) می‌باشد.

مقاله حاضر مدل خطای اندازه‌گیری را به صورت کامل و سه مدل مهم از مدل‌های خطای اندازه‌گیری را بیان کرده و به برآورد پارامتر شیب در این مدل‌ها می‌پردازد.

[۱] Fuller

[۲] Measurement Error (ME) Models

۲ مدل خطای اندازه‌گیری

مدل رگرسیون خطی ساده ی استاندارد با یک متغیر توضیحی را در نظر بگیرید:

$$y = \beta_0 + \beta_1 \gamma + \varepsilon, \qquad (۱)$$

که در آن متغیر مستقل γ، ثابت و یا تصادفی بوده و خطای ε دارای میانگین صفر و ناهمبسته با γ می‌باشد. با توجه به مشاهدات مستقل

$$\begin{pmatrix} \gamma_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} \gamma_n \\ y_n \end{pmatrix},$$

عرض از مبدأ مجهول β_0 و شیب β_1 معمولاً با استفاده از روش کمترین مربعات و یا برخی روش دیگر برآورد می‌شوند[۸].

در یک مدل ME، فرض می‌کنیم که متغیرهای γ و η به صورت زیر به هم مرتبط‌اند:

$$\eta = \beta_0 + \beta_1 \gamma, \qquad (۲)$$

اما این متغیرهای γ و η غیر قابل مشاهده هستند و تنها می‌توانند همراه با خطا مشاهده شوند. بنابراین بجای مشاهده γ و η به صورت مستقیم، مشاهدات زیر را داریم:

$$x = \gamma + \delta, \quad y = \eta + \varepsilon, \qquad (۳)$$

که در اینجا فرض می‌شود متغیر γ و خطاهای δ و ε ناهمبسته هستند[۲].

فرض کنید نمونه‌ای به حجم n برای این مدل به صورت

زیر خواهد بود:

$$\begin{pmatrix} x_۱ \\ y_۱ \end{pmatrix}, ..., \begin{pmatrix} x_n \\ y_n \end{pmatrix}. \qquad (۴)$$

از آنجا که هر دو متغیر γ و η همراه با خطا اندازه‌گیری می‌شوند، این مدل را مدل خطای اندازه‌گیری می‌نامند.

در مدل‌های ME ، متغیرهای γ و η گاهی اوقات متغیرهای پنهان[۳] در برخی از زمینه‌های کاربرد نامیده می‌شوند. زمینه‌هایی که در آن متغیر γ همراه با خطای اندازه‌گیری است شاید بیشتر از زمانی که در آن‌ها γ به دقت اندازه‌گیری می‌شود، مورد استفاده قرار می‌گیرند. اکثر متغیرهای پزشکی مانند فشارخون، ضربان قلب، درجه حرارت و شیمی خون همراه با خطای قابل توجهی اندازه‌گیری می‌شوند. همچنین متغیرهای کشاورزی مانند بارش باران، نیتروژن خاک، درجه آلودگی آفت، تخصیص وسعت زمین کشاورزی به محصول و مانند آن نیز نمی‌توانند به دقت اندازه‌گیری شوند. در علوم مدیریت، علوم اجتماعی و تقریبا هر رشته نزدیک به این رشته‌ها، بسیاری از متغیرها تنها می‌توانند همراه با خطا اندازه‌گیری شوند. به این مدل، مدل ME استاندارد نیز می‌گویند چرا که جمله خطا در معادله فوق وجود ندارد. در اینجا لازم می‌دانیم تعریفی از مدل ME همراه با خطا ارائه نماییم.

۱.۲ مدل خطای معادلات

مدل‌های خطای معادلات[۴] تغییر شکلی از مدل‌های ME استاندارد می‌باشند که معمولاً به عنوان مثال در اندازه‌گیری تفسیر و بررسی پدیده‌های اقتصادی مورد استفاده قرار می‌گیرند.

اکنون وضعیتی را در نظر بگیرید که در آن ارزش‌های واقعی γ و η به دلیل وجود خطای q در معادله به صورت کاملاً خطی به هم مرتبط نیستند. به عبارت دیگر رابطه ی (۲) را برای نمونه‌ای به حجم n بدین گونه خواهیم داشت:

$$\eta_i = \beta_۰ + \beta_۱\gamma_i + q_i, \qquad (۵)$$

که q_i ها متغیرهای تصادفی مستقل دارای میانگین صفر و واریانس مثبت $\sigma_q^۲$ می‌باشند و q_i و γ_j به ازای تمامی i و j ها از هم مستقل می‌باشند.
بنابراین فرض می‌شود که $(\delta_i, \varepsilon_i)$ و (q_j, γ_j) به ازای همه i و j ها مستقل هستند. مدل های (۵) و (۳) مدل خطای معادلات یا مدل خطا در معادلات نامیده می‌شود. بیشتر مؤلفان مدل خطای معادلات را بدین صورت می‌نویسند:

$$x_i = \gamma_i + \delta_i, \quad y_i = \beta_۰ + \beta_۱\gamma_i + e_i, \qquad (۶)$$
$$i = ۱, ۲, ..., n,$$

جمله e_i که خطای اندازه‌گیری نامیده می‌شود، شامل خطای متغیر مستقل بعلاوه خطای معادله به صورت $e_i = \varepsilon_i + q_i$ می‌باشد. بنابراین

$$\begin{aligned} Cov\left(\begin{pmatrix} \delta \\ e \end{pmatrix}\right) &= Cov\left(\begin{pmatrix} \delta \\ \varepsilon \end{pmatrix}\right) + \begin{pmatrix} ۰ & ۰ \\ ۰ & \sigma_q^۲ \end{pmatrix} \\ &= \Omega + \begin{pmatrix} ۰ & ۰ \\ ۰ & \sigma_q^۲ \end{pmatrix}. \end{aligned}$$

[۳] Latent Variables
[۴] Eqnarray Error Models

مدل خطای معادله تقریباً مانند مدل های (۲) و (۳) به نظر می‌رسد که تنها تفاوت آن در نمایش q_i است.

۳ سه مدل متناظر در مدل خطای اندازه‌گیری

مدل های ME خطی یک متغیره (۲) و (۳) را می‌توان بر اساس نمونه‌ای به حجم n به شرح زیر فرمول‌بندی کرد. متغیرهای غیر قابل مشاهده (γ_i, η_i) بصورت زیر به هم مرتبط‌اند:

$$\eta_i = \beta_0 + \beta_1 \gamma_i \quad , \quad i = 1, 2, ..., n, \qquad (7)$$

در این صورت مشاهدات (x_i, y_i)، $i = 1, 2, ..., n$، متغیرهای واقعی همراه با خطاهای افزودنی $(\delta_i, \varepsilon_i)$ بصورت زیر هستند:

$$x_i = \gamma_i + \delta_i, \; y_i = \eta_i + \varepsilon_i, \quad i = 1, 2, ..., n. \qquad (8)$$

اکنون فرض می‌کنیم که $\delta_1, \delta_2, ..., \delta_n$ و $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ همگی واریانس محدود دارند و ناهمبسته هستند و (بدون از دست دادن کلیت) میانگین صفر دارند. نتایج زیر به ازاء هر i و j، $i \neq j$ برقرارمی‌باشند:

$$\begin{aligned} E(\delta_i) &= E(\varepsilon_i) = 0, \\ Var(\delta_i) &= \sigma_\delta^2, \quad Var(\varepsilon_i) = \sigma_\varepsilon^2, \\ Cov(\delta_i, \delta_j) &= Cov(\varepsilon_i, \varepsilon_j) = 0, \\ Cov(\delta_i, \varepsilon_i) &= 0. \end{aligned} \qquad (9)$$

در بیشتر نتایج نظری نیاز است که خطاها مستقل و نه فقط ناهمبسته باشند. غالباً فرض بر این است که خطاهای $(\delta_i, \varepsilon_i)$ بصورت نرمال توزیع شده‌اند[۳]. بر این اساس فرض ناهمبستگی به طور خودکار به فرض استقلال تبدیل شود. فرض نرمال می‌تواند تأثیرات مهمی در جهت شناسایی مدل داشته باشد. بدین معنی که اگر در مدل‌های ME، γ نرمال نباشد، از چندین روش برآورد به عنوان مثال برآورد بر پایه گشتاورهای مرتبه بالا، استفاده می‌شود[۱۰]. چنانچه مدل غیر نرمال نزدیک به یک مدل نرمال باشد، این روش‌ها بی‌اساس هستند.

فرض محدود کننده خطاهای δ_i و ε_i این است که آن‌ها هر کدام میانگین‌ها و واریانس‌های مشترک دارند. اگر آن‌ها با میانگین‌های مشترک صفر نباشند (یعنی خطاهای سیستماتیک وجود دارند)، که این خطاها می‌توانند در β_0 جذب شوند؛ بنابراین فرض خطاها با میانگین صفر، محدود کننده نیست. در صورت چنین پیامدی ممکن است یک مدل بدون عرض از مبدأ به یک مدل با عرض از مبدأ تبدیل شود.

از فرم‌های (۲) و (۳) بسته به مفروضات در مورد γ، سه مدل جداگانه وجود دارد که در ادامه به آن ها می پردازیم.

۱.۳ مدل تابعی

اگر γ_i ها ثابت‌های ناشناخته باشند در نتیجه مدل به عنوان مدل تابعی[5] شناخته می‌شود. که در این حالت داریم:

$$Var(\gamma_i) = 0.$$

[5]Functional Model

۲.۳ مدل ساختاری

اگر γ_i ها متغیرهای تصادفی دارای توزیع یکسان و مستقل از(نه فقط ناهمبسته با) خطاها باشند، مدل به عنوان یک مدل ساختاری[۶] شناخته می‌شود.
برای مدل ساختاری داریم:

$$E(\gamma_i)=\mu, Var(\gamma_i)=\sigma^2, \quad i=1,2,...,n.$$

۳.۳ مدل وراساختاری

در مدل سوم که مدل وراساختاری[۷] نامیده می‌شود، فرض بر این است که γ_i ها متغیرهای تصادفی مستقل اما بطور یکسان توزیع نشده باشند، همراه با میانگین‌های احتمالا متفاوت μ_i و واریانس مشترک σ^2. یعنی

$$E(\gamma_i)=\mu_i, Var(\gamma_i)=\sigma^2, \quad i=1,2,...,n.$$

مدل‌های وراساختاری حالت کلی از مدل‌های تابعی و ساختاری هستند:
اگر $M = \mu_1 = \mu_2 = ... = \mu_n$، در نتیجه مدل وراساختاری به مدل ساختاری و در صورتی که اگر $\sigma^2 = 0$، پس از آن مدل وراساختاری به مدل تابعی کاهش می‌یابد[۱، ۴، ۱۰، ۱۱].

در ادامه به چند مثال در برخی از زمینه‌های کاربرد مدل‌های ME اشاره می‌شود.

مثال ۱.۳. *قانون Boyl مربوط به فشار (P) و حجم (V) برای گسترش بیدررو از یک گاز به صورت $PV^j = C$ است، که در آن j و C پارامترهای برآوردشده هستند. با گرفتن لگاریتم از هر دو طرف یعنی $\eta=\ln(P)$, $\gamma=\ln(V)$ و $\beta_1=-j$، $\beta_0=\ln(C)$ ، فرم $\eta=\beta_0+\beta_1\gamma$ را داریم. قطعا آزمایش انجام شده شامل قانون Boyl است که در آن هر دو متغیر، تنها می‌توانند با خطا اندازه‌گیری شوند. بنابراین یک مدل ME خطی در لگاریتم از متغیرها، مناسب خواهد بود. توجه داشته باشید که مقدار خطای افزودنی برای مدل ME در متغیرهای لگاریتم، تبدیل به خطای ضرب شده در متغیرهای اصلی می‌شود.*

مثال ۲.۳. *فرض کنید خواستار مدلی باشیم که نرخ جرم و جنایت (η) به عنوان تابعی از متوسط درآمد خانواده (γ) باشد. اگر کسی به طور تصادفی اطلاعاتی جمع‌آوری کند آنگاه هر دو متغیر: نرخ جرم و جنایت و متوسط درآمد خانواده، تصادفی هستند و تنها می‌توانند با خطا اندازه‌گیری شوند. از آنجا که γ به طور تصادفی انتخاب شده است، مدل ساختاری مناسب خواهد بود.*

مثال ۳.۳. *مدل تکرار می‌تواند در یک طرح آزمایشی کشاورزی مورد استفاده قرار گیرد، که در آن n قطعه زمین که هر کدام با میزان مختلفی از کود مخلوط شده‌اند و r گیاه که از هر قطعه به طور تصادفی انتخاب شده‌اند، وجود دارد. اگر سطح کود γ و برخی ویژگی‌های گیاه η هر دو با خطا اندازه‌گیری شوند و اگر سطوح کود توسط آزمایشگر تعیین شوند، آنگاه مدل تابعی نمایان می‌شود.*

مثال ۴.۳. *مدل ساختاری می‌تواند در آزمایش کالیبراسیون به وجود آید. یک تولید کننده مواد غذایی بسته‌بندی شده به تازگی می‌خواهد برای اندازه‌گیری رطوبت در یک محصول، رطوبت سنج را بر روی خط در دسترس قرار دهد. تولید کننده خواستار کالیبره کردن این دستگاه جدید با رایج‌ترین روش پذیرفته شده توده هوای اجباری است. نمونه‌ها به وسیله نمونه‌برداری به*

[۶] Structural Model
[۷] Ultrastructural Model

طور تصادفی از مواد تولید شده خط تولید در یک دوره از زمان بدست آمده‌اند. رطوبت در هر نمونه توسط هر دو روش توده هوای اجباری (γ) و سنجش روی خط (η) اندازه‌گیری شده است. هر دو روش اندازه‌گیری رطوبت در معرض خطا هستند. به نظر می‌رسد یک مدل ساختاری به خوبی ارتباط بین دو اندازه‌گیری را گزارش می‌کند[۲].

۴ برآورد پارامتر شیب به روش کمترین مربعات تعمیم‌یافته در مدل خطای اندازه‌گیری

برآوردگر کمترین مربعات، مجموع مربعات از فواصل عمودی را کمینه می‌کند، بدین معنا که تنها خطاها را در برازاندن متغیر وابستهٔ y به حداقل می‌رساند. از این مطلب می‌توان نتیجه گرفت که کمترین مربعات برای مدل‌های خطای اندازه‌گیری (ME) مناسب نمی‌باشند، چرا که مؤلفه خطا در اندازه‌گیری x به حساب نمی‌آیند. هم‌چنین روشن است که برآوردگر کمترین مربعات برای مدل ME اریب است چرا که در مدل رگرسیون معمولی $y = \beta_0 + \beta_1 x + \varepsilon_i$ خطاهای ε_i همراه با متغیرهای مستقل x_i هستند و از اینرو برآوردگر کمترین مربعات برای مدل‌های ME اریب است[۱۲].

یک راه خوب برای پیش‌بینی کمترین مربعات تعمیم‌یافته[۸] از طریق تفسیر هندسی زیر است. کمترین مربعات معمولی، مجموع مربعات فاصله عمودی را به حداقل می‌رساند، چرا که تنها y ها با خطا اندازه‌گیری می‌شوند[۷، ۹]. ولی در مدل‌های خطای اندازه‌گیری، هر دو خطاهای اندازه‌گیری باید در نظر گرفته شوند. چرا که خطاها مستقل هستند، بر همین اساس مجموع مربعات زیر را مینیمم می‌کنیم:

$$\sum \frac{(x_i - \gamma_i)^2}{\sigma_\delta^2} + \sum \frac{(y_i - \beta_0 - \beta_1 \gamma_i)^2}{\sigma_\varepsilon^2} = \sum \frac{\delta_i^2}{\sigma_\delta^2} + \sum \frac{\varepsilon_i^2}{\sigma_\varepsilon^2}. \qquad (10)$$

اگر δ_i و ε_i دارای واریانس یکسان باشند، لازم است که مجموع مربعات زیر را به حداقل برسانیم:

$$\sum (x_i - \gamma_i)^2 + \sum (y_i - \beta_0 - \beta_1 \gamma_i)^2,$$

یعنی،

$$\sum \delta_i^2 + \sum \varepsilon_i^2 = \sum_{i=1}^{n} (\delta_i^2 + \varepsilon_i^2),$$

را مینیمم نماید. اما کمیت $\delta_i^2 + \varepsilon_i^2$ به نوعی فاصله نقطه (x_i, y_i) را از خط به معادله $\eta = \beta_0 + \beta_1 \gamma$ توصیف می‌کند. زیرا با توجه به شکل (۱) اگر نقطه A را معرف نقطه (x_i, y_i) در نظر بگیریم و AB فاصله عمودی نقطه A از خط $\eta - \beta_0 - \beta_1 \gamma = 0$ باشد و نقطه B (پای عمود) روی خط مذکور و در نتیجه معرف نقطه (γ_i, η_i) خواهد بود. از این رو در مثلث قائم الزاویه ABC که در C قائم است، داریم:

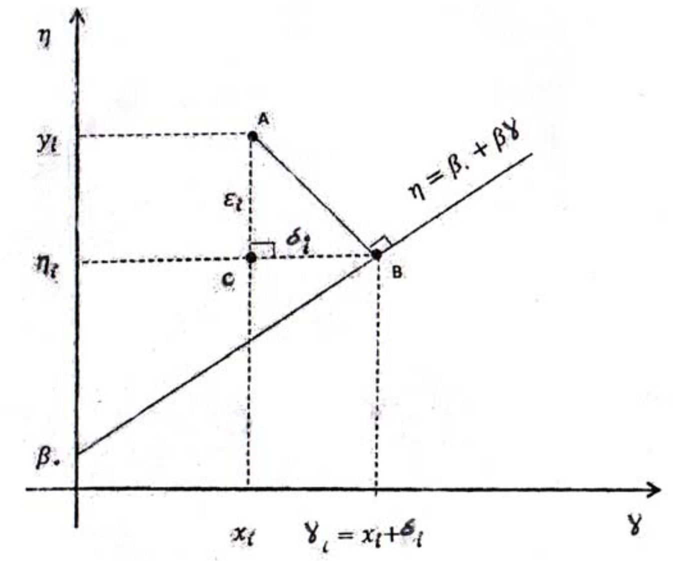
$$AC = \mid y_i - \eta_i \mid = \mid \varepsilon_i \mid, \quad BC = \mid x_i - \gamma_i \mid = \mid \delta_i \mid,$$

و بر اساس قضیه فیثاغورس داریم:

$$(AB) = (AC)^2 + (BC)^2 = \delta_i^2 + \varepsilon_i^2.$$

در این صورت لازم است β_0 و β_1 آن چنان برآورد شوند

[۸] Generalized Least Squares



شکل ۱: فاصله رگرسیونی متعامد

که تا حد ممکن خط مذکور به نقاط به مختصات (x_i, y_i) نزدیک شود. یکی از این روش‌ها این است که مجموع فواصل همه نقاط نظیر A از خط مینیمم شود. با توجه به ریاضی مقدماتی، فاصله عمودی نقطه (x_i, y_i) از خطی به معادله $ax + by + c = 0$ عبارت است از:

$$AB = \frac{| ax_i + by_i + c |}{\sqrt{a^2 + b^2}}.$$

در نتیجه فاصله عمودی A از خطی به معادله $\eta - \beta_0 - \beta_1\gamma = 0$ عبارت است از:

$$AB = \frac{| y_i - \beta_0 - \beta_1 x_i |}{\sqrt{1 + \beta_1^2}}.$$

لذا داریم:

$$(AB)^2 = \delta_i^2 + \varepsilon_i^2 = \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{1 + \beta_1^2}.$$

بنابراین می توان برآوردهای کمترین مربعات تعمیم‌یافته β_0 و β_1 را از طریق مینیمم کردن تابع زیر بدست آورد:

$$\begin{aligned}(\tilde{\beta_0}, \tilde{\beta_1}) &= arg\ min_{(\beta_0, \beta_1)} \\ &\times \sum_{i=1}^{n} \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sqrt{1 + \beta_1^2}} \right)^2, \quad (11)\end{aligned}$$

که این همان رگرسیون متعامد کلاسیک[9] است.

حال در صورت نابرابری واریانس‌های خطا، مدل های (۵) و (۶) را در نظر می‌گیریم و می‌نویسیم:

$$y_i = \beta_0 - \beta_1 x_i + (\varepsilon_i - \beta_1\delta_i),\ i = 1, 2, ..., n. \quad (12)$$

اگر $\psi_i = \varepsilon_i - \beta_1\delta_i$ ،$i = 1, 2, ..., n$، لازم است که

[9] Classical Orthogonal Regression

مانده‌ها، مستقل و عینا با میانگین صفر و واریانس

$$\sigma_{\psi}^{2}=\sigma_{\varepsilon}^{2}-2\beta_{1}\Sigma_{\delta\varepsilon}+\beta_{1}\Sigma_{\delta\delta}\beta_{1}',$$

توزیع شده باشند، بدون توجه به اینکه مدل تابعی، ساختاری و وراساختاری است. برآوردگر کمترین مربعات تعمیم‌یافتهٔ β_0 و β_1 با به حداقل رساندن، رابطه زیر بدست می‌آیند:

$$\begin{aligned}(\tilde{\beta_0},\tilde{\beta_1}) &= arg\ min_{(\beta_0,\beta_1)} n^{-1}\\ &\times \sum \frac{(y_i-\beta_0-\beta_1 x_i)^2}{\sigma_{\psi}^{2}}. \quad (13)\end{aligned}$$

مشاهده می‌شود زمانی‌که واریانس خطاها ناهمسان است، اصولا ماتریس کوواریانس خطاها معلوم باشد و یا لااقل چند درایه‌ای از آن را معلوم در نظر گرفت که به کمک آن بتوانند برآوردگرهای کمترین مربعات تعمیم‌یافته را بیابند.

روش دیگری نیز وجود دارد که بتواند در شرایط وجود واریانس ناهمسانی یا وابستگی پیاپی جملات خطا، مدل رگرسیونی را تخمین بزند.

مدل رگرسیونی زیر را در نظر بگیرید:

$$y=X\beta+\varepsilon,\ E(\varepsilon\varepsilon')=\Omega. \quad (14)$$

با این شرط که Ω یا همان ماتریس کوواریانس جملات خطا یک ماتریس معین‌مثبت است. بر این اساس همان‌طور که می‌دانیم معادلهٔ (۱۴) یک مدل رگرسیون خطی ساده با جملات خطای دارای واریانس یکسان و هم چنین غیر وابسته است. در صورتی که Ω قطری با مقادیر متفاوت روی قطر اصلی باشد، جملات خطا به هم وابسته نیستند اما شرط واریانس همسانی بر قرار نخواهد بود. و در صورتی که ماتریس Ω قطری نباشد، عناصر خطا به هم وابسته خواهند بود. در اقتصاد سنجی قطری نبودن ماتریس عناصر خطا غالباً در هنگام کار بر روی داده‌های سری زمانی به وجود می‌آید. در چنین حالتی عناصر خطا در داده‌هایی که به لحاظ زمانی به هم نزدیک‌ترند وابستگی بالاتری دارند (برای مشاهده مثال کاربردی در رابطه با داده‌های سری زمانی به فاکس و ویزبرگ [۵] مراجعه شود). در ادامه با استفاده از مدل کمترین مربعات تعمیم یافته نشان می‌دهیم که چگونه می‌توان یک برآوردگر کارا برای تخمین بردار β در معادلهٔ (۱۴) به دست آورد که شرایط قضیه گاووس–مارکوف[10] را ارضا نماید.

در نتیجه برای به دست آوردن برآوردگر کارا برای بردار پارامترهای β در یک مدل رگرسیون خطی، از مدل تبدیل یافته‌ای[11] برای رسیدن به این هدف استفاده می‌شود.

در مدل رگرسیون خطی کلاسیک $\boldsymbol{y}=\boldsymbol{X\beta}+\varepsilon$،

$$Var(\varepsilon\mid \boldsymbol{X})=\sigma^2 I_n\ ,E(\varepsilon\mid \boldsymbol{X})=0,$$

برقرار و

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}, \quad (15)$$

بهترین برآوردگر نااریب خطی است. با فرض نرمال بودن خطا با میانگین صفر. اکنون فرض کنید که تمامی پیش‌فرض‌های کلاسیک برقرار باشند به استثنای واریانس خطا یعنی داشته باشیم:

$Var(\varepsilon\mid \boldsymbol{X})=\sigma^2\Omega$، که اگر Ω یک ماتریس $n\times n$ متقارن، بالا مثلثی و معین مثبت باشد، آنگاه یک ماتریس ناتکین $\boldsymbol{C}_{(n\times n)}$ وجود دارد به طوری که $\Omega^{-1}=\boldsymbol{C'C}$.

در نتیجه این تبدیل با استفاده از ماتریس C که غالباً

[10] Gauss-Markov Theorem

[11] Transformed Model

یک ماتریس بالا مثلثی است، صورت می‌پذیرد. اکنون مدل تبدیل‌یافته زیر را در نظر می‌گیریم:

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \varepsilon^*,$$

که $\boldsymbol{Y}^* = \boldsymbol{CY}$ ، $\boldsymbol{X}^* = \boldsymbol{CX}$ و $\varepsilon^* = \boldsymbol{C}\varepsilon$ و در نتیجه مدل تبدیل‌یافته به طور بدیهی خطی است. داریم:

$$\begin{aligned} E(\varepsilon^* \mid \boldsymbol{X}) &= \boldsymbol{C}E(\varepsilon \mid \boldsymbol{X}) = ۰, \\ Var(\varepsilon^* \mid \boldsymbol{X}) &= Var(\boldsymbol{C}\varepsilon \mid \boldsymbol{X}) \\ &= E(\boldsymbol{C}\varepsilon\varepsilon'\boldsymbol{C}' \mid \boldsymbol{X}) \\ &= \boldsymbol{C}E(\varepsilon\varepsilon' \mid \boldsymbol{X})\boldsymbol{C}' \\ &= \boldsymbol{C}\sigma^۲\Omega\boldsymbol{C}' = \sigma^۲\boldsymbol{C}(\Omega^{-۱})^{-۱}\boldsymbol{C}' \\ &= \sigma^۲\boldsymbol{C}\left[(\boldsymbol{C}'\boldsymbol{C})^{-۱}\right]^{-۱}\boldsymbol{C} = \sigma^۲ I_n. \end{aligned}$$

بنابراین از آنجا که مدل تبدیل‌یافته تمامی فرض‌های کلاسیک را حفظ کرده است در نتیجه قضیه گاووس-مارکوف برای مدل تبدیل‌یافته قابل استفاده است، از اینرو بهترین برآوردگر نااریب خطی به صورت زیر است:

$$\hat{\beta} = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-۱}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^*. \qquad (۱۶)$$

این برآوردگر کمترین مربعات تعمیم‌یافته است. به طور کلی می‌توان گفت برای مدل رگرسیون خطی چندگانه

$$\boldsymbol{y} = \boldsymbol{X\beta} + \varepsilon,$$

که در آن $\boldsymbol{y}$ بردار n بعدی متغیر پاسخ و $\boldsymbol{X}$ ماتریس متغیرهای توضیحی و ε بردار خطا است. در حالت کلی ε دارای ماتریس واریانس کوواریانس Σ و Σ وارون‌پذیر است، یک برآوردگر برای بردار پارامتر β با استفاده از کمترین مربعات تعمیم‌یافته به صورت زیر می‌باشد:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\Sigma^{-۱}\boldsymbol{X})\boldsymbol{X}'\Sigma^{-۱}\boldsymbol{y}. \qquad (۱۷)$$

در حقیقت برآوردگر کمترین مربعات تعمیم‌یافته در حالت یک متغیره، برای مدل بدون معادله خطا یا ماتریس کوواریانس خطای تا چند عددی شناخته شده کار می‌کند.

۵ بحث و نتیجه‌گیری

همانطور که می‌دانیم هنگامی که متغیرهای رگرسیونی در معرض خطا اندازه‌گیری شوند، ناگزیر به استفاده از مدل‌های خطای اندازه‌گیری هستیم. در چنین مواردی چون روش کمترین مربعات معمول در برآورد ضرایب مدل منجر به یافتن برآوردهایی اریب می‌شود و با توجه به اهمیت مبحث نااریبی در بحث برآوردگرها، در این نوشتار راهکاری به روش کمترین مربعات تعمیم‌یافته برای تصحیح نااریبی برآوردگرهای کمترین مربعات ضرایب رگرسیونی در مدل‌های خطای اندازه‌گیری معرفی شدند.

مراجع

[1] Brown, G. H. (1978). Generalized least squares applied to the linear ultrastructural models. Biometrika, 65, 441-444.

[2] Cheng, C. and Van Ness, J. W. (1999). Statistical Regression with Measurement Errors. Wiley, New York.

[3] Cox, N. R. (1976). The linear structural relation for several groups of data. Biometrika, 63, 231-237.

[4] Dolby, G. R. (1976). The ultrastructural model: a synthesis of the functional and structural relations. Biometrika, 63, 39-50.

[5] Fox, J. and Weisberg, S. (2010). Time-series regression and generalized least squares in R, An Appendix to An R Companion to Applied Regression, Second Edition.

[6] Fuller, W. A. (1987). Measurement Error Models. Wiley, New York.

[7] Hasabelnaby, N. A., Ware, J.H. and Fuller, W.A. (1989). Indoor air pollution and pulmonary performance: investigating errors in exposure assessment (with comments). Statist. Med., 8(9), 1109-1126.

[8] Kuan, C. M. (2004). The Method of Generalized Least Squares. Wiley, New York.

[9] Richardson, D.W. and Wu, D. (1970). Least squares and grouping method estimators in the errors in variabes models. J. Amer. Statist. Assoc., 65, 724-748.

[10] Srivastava, A. K. and Shalabh. (1997). Consistent estimation for the non-normal ultrastructural model. Statist. Probab. Lett., 34(1), 67-73.

[11] Srivastava, A. K. and Shalabh. (1997). Improved estimation of slope parameter in a linear ultrastructural model when measurement errors are not necessarily normal. J. Econometrics, 78, 153-157.

[12] Wang, C. Y. (1993). Alternative covariance estimates in a replicated measurement error model with correlated heteroscedastic errors. Comm. Statist. Theory Methods, 22, 1819-1828.