سیزدهمین کنفرانس آمار ایران
۴-۲ شهریور ۱۳۹۵

براورد رگرسیون ریج برای مدل خطای اندازه گیری در حضور متغیر ابزاری

دکتر ایوب شیخی [1]، مهدیه شمسی نژاد [۲]

[1]دانشگاه شهید باهنر کرمان

[2]دانشگاه شهید باهنر کرمان

چکیده: این مقاله براورد پارامترهای مدل خطای اندازه گیری که در آن ماتریس کواریانس پارامترهای رگرسیونی شرایط مساعدی ندارند را ارائه می کند. در بسیاری از موارد عملی مشاهدات نه تنها با خطا اندازه گیری شده اند بلکه روابط خطی با خودشان نیز دارند، بنابراین مشکل همخطی چندگانه در آن داده ها نمایان می شود. وقتی روش کمترین مربعات در مورد داده های غیرمتعامد بکار گرفته میشود. معمولا برای ضرایب رگرسیون براوردهای خیلی ضعیف بدست می آید. رگرسیون ریج معرفی شده توسط هورل و کنارد (۱۹۷۰) را برای این شرایط و برای تخمین پارامترهای مدل خطای اندازه گیری در نظر می گیریم. فرض خطای اندازه گیری کلاسیک مدعی است که خطای اندازه گیری در هر متغیری از مشاهدات، مستقل از همه متغیرهای واقعی مشاهده شده ی دیگر است. حال اگر فرضیه بالا برای مدل خطای اندازه گیری برقرار نباشد، نشان می دهیم براوردگر متغیر ابزاری برای مدل سازگار خواهد بود. مدل خطای اندازه گیری می تواند حالت خاصی از مدلهایی با رگرسورهای درون زا در نظر گرفته شود، از این رو روش متغیرهای ابزاری معمول ترین روش بدست آوردن براوردهای سازگار مدل رگرسیون خطی با خطای اندازه گیری کلاسیک است.

واژه‌های کلیدی: همخطی چندگانه، رگرسیون ریج، خطای اندازه گیری، متغیر ابزاری و براوردگرهای انقباضی.

۱ مقدمه

این مقاله براورد پارامترهای مدل خطای اندازه گیری که در آن ماتریس کواریانس پارامترهای رگرسیونی شرایط مساعدی ندارند را ارائه می کند. رگرسیون ریج معرفی شده توسط هورل و کنارد (۱۹۷۰) را برای این شرایط و برای تخمین پارامتر های مدل خطای اندازه گیری وقتی که پارامترها به یک زیر فضای خطی تعلق دارند در نظر می گیریم.

یکی از فرضیات اساسی در آنالیز رگرسیون خطی آن است که همه متغیر های توضیحی به طور خطی مستقلند و رتبه ماتریس مشاهدات روی همه متغیر های توضیحی پر رتبه است. وقتی این فرضیه نقض می شود مشکل هم خطی چند گانه در داده ها داریم. واریانس براوردگر کمترین مربعات معمولی ضرایب رگرسیون به خاطر هم خطی چندگانه مینیمم نیست. مشکل هم خطی چند گانه توجه بسیاری از محققان را به خود جلب کرده است. سیلوی(۱۹۶۹)،بلسلی(۱۹۹۱)،که و همکاران(۲۰۰۴) و رائو و همکاران(۲۰۰۸).

[۲]مهدیه شمسی نژاد : shamsi.m@math.uk.ac.ir

به دست آوردن براوردگرها برای داده های هم خطی چندگانه یکی از اهداف مقاله است که روشهای متعددی برای ان ارائه شده است. که در میان آنها روش براورد رگرسیون ریج معرفی شده توسط هورل و کنارد محبوب ترین روش در میان محققان است. براوردگر ریج تحت فرض خطای تصادفی نرمال در مدل رگرسیونی توسط گیبونز(۱۹۸۱)،سارکر(۱۹۹۲)،صالح و همکاران(۱۹۹۳)،گروبر(۱۹۹۸)،مالدوز(۱۹۹۹)، سان و همکاران(۱۹۹۹)،اینوئو(۲۰۰۱)،کیبریا و همکاران(۲۰۰۳و۲۰۰۴)، آرشی و همکاران(۲۰۱۰)،حسن زاده وهمکاران(۲۰۱۱)،آرشی و همکاران(۲۰۱۲)،بشتیان و همکاران(۲۰۱۲). ارائه شده است. جرئیات و توسعه روش های دیگر مرتبط با رگرسیون ریج در حوصله این مقاله نیست.

یکی دیگر از فرضهای اساسی در تمام تجزیه و تحلیل های آماری آن است که همه مشاهدات به درستی مشاهده شده باشد این فرض اغلب در واقعیت نقض می شود در نتیجه خطای اندازه گیری در داده ها اتفاق می افتد. در این مواقع مشاهدات به درستی مشاهده نشده و با خطای اندازه گیری آمیخته شده اند. ابزار آماری رایج اعتبار خود را هنگامی که داده ها با خطا اندازه گیری شده اند از دست می دهند. فولر(۱۹۷۸) و چانگ و همکاران(۱۹۹۹)

در زمینه مدل های رگرسیون خطی براوردگر کمترین مربعات معمولی ضرایب رگرسیون در حضور خطای اندازه گیری در داده ها بخوبی سازگار نیستند. مسئله مهم در مدل خطای اندازه گیری پیداکردن براوردگر سازگار پارامتر است. براوردگر های سازگار ضرایب رگرسیون با استفاده از برخی اطلاعات اضافی خارج از نمونه بدست می آید.

درزمینه مدل های رگرسیون خطی چندگانه اطلاعات اضافی را در قالب ماتریس کواریانس شناخته شده از خطای اندازه گیری مرتبط با متغیر توضیحی و همچنین ماتریس شناخته شده ی نسبت قابل اطمینان از متغیر های توضیحی[۱] فولر و همکاران(۱۹۷۸)، گلسر(۱۹۹۲)، کیم و همکاران(۲۰۰۲و۲۰۰۳)، صالح(۲۰۱۰)، سان و همکاران(۲۰۱۰)،جورکوف و همکاران(۲۰۱۰) و شلاب و همکاران(۲۰۱۱) به کار می برند.

۲ توصیف مدل

در بسیاری از موارد عملی مشاهدات نه تنها با خطا اندازه گیری شده اند بلکه روابط خطی بین خودشان نیز دارند. بنابراین مشکل هم خطی چندگانه در داده هایی که با خطا اندازه گیری شده اند نمایان می شوند. مسئله مهم چگونگی به دست آوردن براوردگر های ضرایب رگرسیون تحت این شرایط است. یک ایده ساده استفاده از براورد رگرسیون ریج در این شرایط است. به عبارت دیگر رگرسیون ریج معرفی شده توسط هورل و کنارد مشکل هم خطی چندگانه برای براورد پارامتر های رگرسیون را در مدل خطای اندازه گیری حل می کند.

مدل رگرسیون خطی چندگانه با خطای اندازه گیری را در نظر بگیرید

$$y_t = \beta_0 + x_t'\boldsymbol{\beta} + e_t, \quad \mathbf{X}_t = \mathbf{x}_t + \mathbf{u}_t, \quad t = 1, 2, \ldots, n$$

که در آن β_0 عرض از مبدا خط رگرسیونی است و $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ بردار ضرایب رگرسیون $p \times 1$ است، همچنین $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{pt})'$ یک بردار $p \times 1$ است که $t-$امین مشاهده درست اما غیرقابل مشاهده را نشان می دهد . $p-$متغیر توضیحی که مشاهده شده اند بصورت بردار $\mathbf{X}_t = (X_{1t}, X_{2t}, \ldots, X_{pt})'$ نمایش داده شده است و بردار $p \times 1$ خطای اندازه گیری $\mathbf{U}_t = (u_{1t}, u_{2t}, \ldots, u_{pt})'$ که u_{it} خطای اندازه گیری در $i-$امین متغیر توضیحی x_{it} است. e_t خطای تصادفی مدل رگرسیون در متغیر مشاهده شده Y_t است.

[۱] Reliability ratio matrix

فرض می کنیم که

$$(\mathbf{x}_t', e_t, \mathbf{u}_t) \sim N_{2p+1}\left\{(\boldsymbol{\mu}_t', 0, \mathbf{0}'), \text{BlockDiag}\left(\Sigma_{xx}, \sigma_{ee}, \Sigma_{uu}\right)\right\}$$

با $\boldsymbol{\mu}_x = (\mu_{x1}, \mu_{x2}, \ldots, \mu_{xp})'$ و σ_{ee} واریانس e_t است.

Σ_{xx} و Σ_{uu} که به ترتیب ماتریس های کواریانس x_t و u_t است.

واضح است که $(y_t, \mathbf{x}_t')'$ توزیع نرمال چندمتغیره $p+1$ متغیره است با بردار میانگین $(\beta_0 + \beta'\mu_x, \mu_x')'$ و ماتریس کواریانس زیر است

$$\begin{pmatrix} \sigma_{ee} + \boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta} & \boldsymbol{\beta}'\Sigma_{xx} \\ \Sigma_{xx}\boldsymbol{\beta} & \Sigma_{xx} + \Sigma_{uu} \end{pmatrix}.$$

پس توزیع شرطی y_t بشرط x_t' برابر است با

$$E(Y_t|X_t) = \gamma_0 + \boldsymbol{\gamma}'\mathbf{X}_t$$

که در آن γ_0 و γ و σ_{zz} بصورت ذیل هستند

$$\gamma_0 = \beta_0 + \boldsymbol{\beta}'(I_p - k_{xx}')\boldsymbol{\mu}_x, \qquad \boldsymbol{\gamma} = k_{xx}\boldsymbol{\beta}, \qquad \boldsymbol{\beta} = k_{xx}^{-1}\boldsymbol{\gamma},$$
$$k_{xx} = \Sigma_{XX}^{-1}\Sigma_{xx} = \left(\Sigma_{xx} + \Sigma_{uu}\right)^{-1}\Sigma_{xx}.$$
$$\Sigma_{zz} = \Sigma_{ee} + \boldsymbol{\beta}'\Sigma_{xx}(I_p - k_{xx}')\boldsymbol{\beta}$$

و k_{xx} ماتریس $p \times p$ از نسبت های قابل اطمینان X است. گلسر(۱۹۹۲)، فولر و همکاران(۱۹۷۸). [4,7]

مشکل اساسی در براورد β تحت فرضیـات مختلف را با براورد مقدماتی β بـفرض دانسـتن Σ_{xx} شروع میکنیم. قرار میدهیم

$$S = \begin{pmatrix} S_{YY} & S_{Yx} \\ S_{xY} & S_{XX} \end{pmatrix}$$

$$S_{YY} = (\mathbf{Y} - \bar{Y}\mathbf{1}_p)'(\mathbf{Y} - \bar{Y}\mathbf{1}_p)$$

$$S_{XX} = (S_{X_iX_i}), S_{X_iX_i} = (\mathbf{x}_i - \bar{X}_i\mathbf{1}_n)'(\mathbf{x}_i - \bar{X}_i\mathbf{1}_n)$$

$$S_{X_iY} = (\mathbf{X}_i - \bar{X}_i\mathbf{1}_n)'(\mathbf{Y}_i - \bar{Y}_i\mathbf{1}_n), S_{XY} = (S_{X_1Y}, S_{X2Y}, \ldots, S_{X_PY})$$

$$\bar{X}_i = \frac{1}{n}\sum_{t=1}^{n} X_{it}, \bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t.$$

واضح است $\frac{1}{n-1}S_{XX}$ براورد ناااریب Σ_{XX} و $\frac{1}{n}S_{XX} \xrightarrow{p} \Sigma_{XX}$ وقتی که $n \to \infty$ میل میکند. در آن $\xrightarrow{p}$ نشان دهنده ی همگرایی در احتمال است. گلسر(۱۹۹۲)

یا [۵] نشان داده که براورد درستنمایی ماکزیممγ_0 و γ و σ_{zz} همانند براورد کمترین مربعات است

$$\tilde{\gamma}_{0n} = \bar{Y} - \tilde{\boldsymbol{\gamma}}_n'\bar{\mathbf{X}}, \qquad \tilde{\boldsymbol{\gamma}}_n = S_{XX}^{-1}S_{XY},$$
$$\tilde{\sigma}_{zz} = \frac{1}{n}(Y - \tilde{\gamma}_{0n}\mathbf{1}_n - \tilde{\boldsymbol{\gamma}}_n'\mathbf{X})'(Y - \tilde{\gamma}_{0n}\mathbf{1}_n - \tilde{\boldsymbol{\gamma}}_n'\mathbf{X})$$

که تولید می کند:

$$\tilde{\sigma}_{ee} = \tilde{\sigma}_{zz} - \tilde{\boldsymbol{\gamma}}_n'k_{xx}^{-1}\Sigma_{uu}\tilde{\boldsymbol{\gamma}}_n \geq 0$$

که درآن Σ_{uu} شناخته شده است و k_{xx} ناشناخته است و بوسیله جایگذاری مداوم $\Sigma_{xx} + \Sigma_{uu}$ و Σ_{xx} توسط براوردگرهای سازگار مربوطه شان بدست می اید و با توجه به اینکه $\frac{1}{n}S_{XX}$ براورد درستنمایی ماکسیمم $\Sigma xx + \Sigma_{uu}$ نیز هست؛ داریم:

$$k_{xx} = {\Sigma_{XX}}^{-1}\Sigma_{xx} = (\Sigma_{xx} + \Sigma_{uu})^{-1}\Sigma_{xx}$$

$$\hat{k}_{xx} = S_{XX}^{-1}\left(S_{XX} - n\Sigma_{uu}\right)$$

که $\hat{k}_{xx} \xrightarrow{p} k_{xx}$ وقتی که $n \to \infty$ میل میکند.

پس براوردهای درستنمایی ماکسیمم پارامترها می شود:

$$\tilde{\beta}_{0n} = \tilde{\gamma}_{0n} + \tilde{\boldsymbol{\beta}}_n'(I_p - \hat{k}_{xx}')\bar{\mathbf{x}}, \qquad \tilde{\boldsymbol{\beta}}_n = \hat{k}_{xx}^{-1}\tilde{\boldsymbol{\gamma}}_n,$$

$$\tilde{\sigma}_{ee} = \tilde{\sigma}_{zz} - \tilde{\boldsymbol{\beta}}_n'\Sigma_{uu}\hat{k}_{xx}\tilde{\boldsymbol{\beta}}_n$$

و سرانجام با کاهش $\tilde{\beta}_{0n}$ به $\bar{Y} - \tilde{\boldsymbol{\beta}}_n'\bar{\mathbf{X}}$ خواهیم داشت:

$$\tilde{\boldsymbol{\beta}}_n = (S_{XX} - n\Sigma uu)^{-1}S_{XY}$$

که باز هم همانند قبل $\tilde{\sigma}_{ee} \geq 0$ می شود.

براوردها بعنوان براوردهای نامتناهی پارامتر β با توجه به قضیه۲فولر(۱۹۷۸) [۴] وقتی که $n \to \infty$ میل می کند قابل تحقیق خواهند بود.

وقتی روش کمترین مربعات در مورد داده های غیر متعامد بکار گرفته می شود. معمولا برای ضرایب رگرسیون براوردهای خیلی ضعیف بدست می آید. که واریانس برآوردهای کمترین مربعات ضرایب رگرسیون در حد قابل توجهی افزایش می یابد . طول بردار براورد کمترین مربعات پارامتر بطور متوسط خیلی زیاد است و این یعنی قدر مطلق براوردهای کمترین مربعات خیلی بزرگ می باشد و لذا خیلی ناپایدار می باشد بدین معنی که با ارائه نمونه های متفاوت اندازه ها و علامت ها در حد قابل توجهی تغییر خواهند کرد که این شرایط نامساعد نمونه ای از هم خطی چندگانه است که باعث می شود یک یا بیشتر مقادیر ویژه کوچک و نزدیک صفر شود.

هنگامی که ارتباط خطی نزدیکی بین متغیرهای رگرسیونی است گفته می شود مسئله هم خطی چندگانه وجود دارد. اگر مرز تعریف x کوچک باشد اضافه کردن یک جمله هم می تواند هم خطی تولید کند. همچنین یک مدل با متغیرهای رگرسیونی بیش از حد مثل داده های پزشکی و ژنتیک که نمونه ها کمتر از متغیرهای رگرسیونی اند یعنی برای هر فرد تعداد زیادی متغیر رگرسیونی جمع آوری می شود که این باعث همخطی میشود.

برای رفع مشکل هم خطی در مدل از براوردگرهای انقباضی نظیر ridge regression و lasso استفاده می شود. که براوردگر ridge بخاطر استفاده از توان دوم خطا در مقابل براوردگر lasso که قدر مطلق خطاست بعلت سادگی ارجحیت دارد .

پیدا کردن مقدار λ در براورد رگرسیون ریج $(\beta = (X'X - \lambda I)^{-1}X'Y)$ بطوری که کاهش واریانس، موثرتر از افزایش اریبی باشد همیشه مورد توجه بوده است. به همین دلیل محققان برای λ براوردهای مختلفی بکار برده اند که تا کنون حدود ۱۶ براورد برای آن برحسب نیاز و دانسته های قبلی معرفی شده اند. هورل و همکاران(۱۹۷۰)

بسیاری از مشاهده های اقتصادی ، پزشکی ، صنعتی و ژنتیک همواره با متغیرهای مشکوک به خطا در اندازه گیری به دست می آیند. مشکل خطای اندازه گیری یکی از مشکلات اساسی در این رشته هاست بخاطر خطای دید و ... حضور خطاهای اندازه گیری باعث

براورد ناسازگار و اریب پارامترها در روش کمترین مربعات کلاسیک می شود و منجربه نتیجه گیری های نادرست در تجزیه و تحلیل ها می شود .

به این دلیل روش های دیگری که اساس آن ها روی خطای متغیرهای خطی است معرفی می شوند . روش های متفاوت برای رفتار خطای اندازه گیری کلاسیک و غیر کلاسیک استفاده شده است. خطای اندازه گیری کلاسیک است اگر متغیرهای واقعی ولی نهفته مستقل باشند و در غیر این صورت غیر کلاسیک است. روشهای مختلف برای مدل خطای متغیر های خطی (EIV) با خطای اندازه گیری کلاسیک در حال حاضر شناخته شده اند. و بطور گسترده مورد استفاده قرار می گیرند. فولر(۱۹۷۸)، کاروولو همکاران(۱۹۹۵)، وانسبیک و همکاران(۲۰۰۰)، بوند و همکاران(۲۰۰۱) و هاسمان(۲۰۰۱).

روش کلی استفاده از مدل EIV که مخفف (Error- In- Variable) است. فرض نرمال چند متغیره است و فرض خطای اندازه گیری کلاسیک مدعی است که خطای اندازه گیری در هر متغیری از مشاهدات، مستقل از همه متغیرهای واقعی مشاهده شده دیگراست .

مدل EIV بفرم زیر است

$$y_t = \beta_0 + x_t'\beta + e_t, \qquad X_t = x_t' + u_t, \qquad t = 1, 2, \ldots, n$$

که در آن x_t' با خطای u_t مشاهده شده است. بعبارت دیگر خطاهای مستقل که در مشاهدات رگرسورها حضور دارند منجر به تضعیف مدل رگرسیون کلاسیک شده و براوردهای ضرایب رگرسیون اریب و ناسازگار می شوند.

خطاهای اندازه گیری اگر متغیر پاسخ را تحت تاثیر قرار دهند تا زمانیکه خطای اندازه گیری ناهمبسته با میانگین صفر و واریانس ثابت باشد هیچ مشکلی پیش نمی آید و با جمله خطای داخل مدل یکی می باشد .تنها نتیجه از حضور خطای اندازه گیری در متغیر پاسخ اینست که آن در خطای مدل رگرسیون پدیدار می شوند.

و حال طبق فرض استفاده از مدل EIV داریم

$$(\mathbf{x}_t, e_t, \mathbf{u}_t) \sim mnv(\boldsymbol{\mu}, 0, \mathbf{0}), \begin{pmatrix} \Sigma_{xx} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_{ee} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{uu} \end{pmatrix}$$

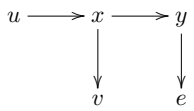
که با این شرط ضرایب رگرسیون سازگار برای پارامترها براورد می شوند [۷].

حال اگر فرضیه بالا برای مدل خطای اندازه گیری برقرار نباشد یعنی

$$(\mathbf{x}_t, e_t, \mathbf{u}_t) \sim mnv(\boldsymbol{\mu}, 0, \mathbf{0}), \begin{pmatrix} \Sigma_{xx} & \mathbf{0} & \Sigma_{xu} \\ \mathbf{0} & \sigma_{ee} & \mathbf{0} \\ \Sigma_{ux} & \mathbf{0} & \Sigma_{uu} \end{pmatrix}$$

نشان می دهیم براوردگر متغیر ابزاری β برای مدل سازگار خواهد بود.

متغیر ابزاری IV را اغلب بصورت دیاگرام زیر برای مدل نمایش می دهند که در آن u از طریق وابستگی به x روی y تاثیر میگذارد. که u در اینجا متغیر ابزاری محسوب می شود [۸].



مدل خطای اندازه گیری می تواند حالت خاصی از مدل هایی با رگرسورهای درون زا در نظر گرفته شود ، ازین رو روش متغیرهای ابزاری معمول ترین روش بدست آوردن برآوردهای سازگار پارامترها در مدل رگرسیون خطی با خطای اندازه گیری مستقل کلاسیک است. بعلاوه می توان از آزمون Hausuam برای بررسی حضور خطاهای اندازه گیری کلاسیک در مدل رگرسیون خطی استفاده کرد [۶]. در عمل متغیر ابزاری IV اغلب از یک متغیر Pronetrue-Error که در معادله دوم مدل EIV است می آید.

و دوباره سعی می کنیم در این مدل جدید براورد ریج رگرسیون را بدست آوریم.

فرض می کنیم که $cov(\mathbf{x}, \mathbf{u}) \neq 0$

$$(\mathbf{x}_t', e_t, \mathbf{u}_t) \sim N_{2p+1} \left\{ (\boldsymbol{\mu}, 0, \mathbf{0}), \begin{pmatrix} \Sigma_{xx} & \mathbf{0} & \Sigma_{xu} \\ \mathbf{0} & \sigma_{ee} & \mathbf{0} \\ \Sigma_{ux} & \mathbf{0} & \Sigma_{uu} \end{pmatrix} \right\}$$

با $\mu_x = (\mu_{x1}, \mu_{x2}, \ldots, \mu_{xp})'$ و واریانس σ_{ee} و واریانس e_t است.

Σ_{xx} و Σ_{uu} به ترتیب ماتریس های کواریانس x_t و u_t است و Σ_{xu} کواریانس بین x_t و u_t است. واضح است که $(y_t, \mathbf{x}_t')'$ توزیع نرمال چندمتغیره $p+1$ متغیره است با بردار میانگین $(\beta_0 + \boldsymbol{\beta}'\mu_x, \mu_x')'$ و ماتریس کواریانس زیر است

$$\begin{pmatrix} \sigma_{ee} + \boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta} & \boldsymbol{\beta}'\Sigma_{xx} + \Sigma_{xu} \\ \Sigma_{xu} + \Sigma_{xx}\boldsymbol{\beta} & \Sigma_{xx} + \Sigma_{uu} + 2\Sigma_{xu} \end{pmatrix}.$$

پس توزیع شرطی y_t بشرط x_t' برابر است با

$$E(Y_t|X_t) = \gamma_0 + \boldsymbol{\gamma}'\mathbf{X}_t$$

که در آن γ_0 و γ و σ_{zz} بصورت ذیل هستند

$$\gamma_0 = \beta_0 + \boldsymbol{\beta}'(I_p - k_{xx}')\boldsymbol{\mu}_x, \qquad \boldsymbol{\gamma} = k_{xx}\boldsymbol{\beta}, \qquad \boldsymbol{\beta} = k_{xx}^{-1}\boldsymbol{\gamma},$$
$$k_{xx} = {\Sigma_{XX}}^{-1}(\Sigma_{xx} + \Sigma_{xu}) = (\Sigma_{xx} + \Sigma_{uu}2\Sigma_{xu})^{-1}(\Sigma_{xx} + \Sigma_{xu}).$$
$$\sigma_{zz} = \sigma_{ee} + \boldsymbol{\beta}'\Sigma_{xx}(I_p - k_{xx}')\boldsymbol{\beta} + \boldsymbol{\beta}'\Sigma_{xu}k_{xx}'\boldsymbol{\beta}$$

و k_{xx} ماتریس $p \times p$ از نسبت های قابل اطمینان X است.گلسر(۱۹۹۲) و فولر و همکاران(۱۹۷۸).

مشکل اساسی در براورد β تحت فرضیـات مختلف را با براورد مقدماتی β بـفرض دانسـتن Σ_{uu} و Σ_{xu} شروع میکنیم. و همانند قبل براورد درستنمایی ماکسیمم پارامترها را بدست می اوریم.

$$\tilde{\gamma}_{0n} = \bar{Y} - \tilde{\boldsymbol{\gamma}}_n'\bar{X}, \qquad \tilde{\boldsymbol{\gamma}}_n = S_{XX}^{-1}S_{XY},$$
$$\tilde{\sigma}_{zz} = \frac{1}{n}(Y - \tilde{\gamma}_{0n}\mathbf{1}_n - \tilde{\boldsymbol{\gamma}}_n'\mathbf{X})'(Y - \tilde{\gamma}_{0n}\mathbf{1}_n - \tilde{\boldsymbol{\gamma}}_n'\mathbf{X})$$

که تولید می کند:

$$\tilde{\sigma}_{ee} = \tilde{\sigma}_{zz} - \tilde{\boldsymbol{\gamma}}_n'k_{xx}^{-1}(\Sigma_{uu} + 2\Sigma_{xu})\tilde{\boldsymbol{\gamma}}_n \geq 0$$

که درآن k_{xx} ناشناخته است و بوسیله جایگذاری مداوم $\Sigma_{xx}+\Sigma_{uu}$ و $\Sigma_{xx}+2\Sigma_{xu}+\Sigma_{uu}$ توسط براوردگرهای سازگار مربوطه شان بدست می اید و با توجه به اینکه $\frac{1}{n}S_{XX}$ براورد درستنمایی ماکسیمم $\Sigma_{xx}+2\Sigma_{xu}+\Sigma_{uu}$ نیز هست؛ داریم:

$$k_{xx}=\Sigma_{XX}{}^{-1}\left(\Sigma_{xx}+\Sigma_{xu}\right)=\left(\Sigma_{xx}+\Sigma_{uu}2\Sigma_{xu}\right)^{-1}\left(\Sigma_{xx}+\Sigma_{xu}\right).$$

$$\hat{k}_{xx}=S_{XX}^{-1}\left(S_{XX}-n\left(\Sigma_{uu}+\Sigma_{xu}\right)\right)$$

که $\hat{k}_{xx}\xrightarrow{p}k_{xx}$ وقتی که $n\to\infty$ میل میکند.

پس براوردهای درستنمایی ماکسیمم پارامترها می شود:

$$\tilde{\beta}_{0n}=\tilde{\gamma}_{0n}+\tilde{\boldsymbol{\beta}}_n'(I_p-\hat{k}_{xx}')\bar{\mathbf{x}},\qquad \tilde{\boldsymbol{\beta}}_n=\hat{k}_{xx}^{-1}\tilde{\boldsymbol{\gamma}}_n,$$

$$\tilde{\sigma}_{ee}=\tilde{\sigma}_{zz}-\tilde{\boldsymbol{\beta}}_n'\left(2\Sigma_{xu}+\Sigma_{uu}\right)\hat{k}_{xx}\tilde{\boldsymbol{\beta}}_n$$

و سرانجام با کاهش $\bar{Y}-\tilde{\boldsymbol{\beta}}_n'\bar{\mathbf{X}}$ به $\tilde{\beta}_{0n}$ خواهیم داشت:

$$\tilde{\boldsymbol{\beta}}_n=\left(S_{XX}-n\left(\Sigma uu+\Sigma_{xu}\right)\right)^{-1}S_{XY}$$

که باز هم همانند قبل $\tilde{\sigma}_{ee}\geq 0$ می شود.

براوردها بعنوان براوردهای نامتناهی پارامتر β با توجه به قضیه ۲ فولر (۱۹۷۸) وقتی که $n\to\infty$ میل می کند قابل تحقیق خواهند بود [۴].

دوباره در اینجا براوردگر رگرسیون ریج را در حضور متغیر ابزاری بدست می اوریم. برای اینکار ابتدا β را از روش کمترین مربعات با دانستن ماتریس قابل اطمینان k_{xx} محاسبه نموده و سپس مینیمم کردن فرم درجه دو، توسط ضرایب لاگرانژ با شرط یا جریمه $\beta'\beta<\lambda$ که شعاع λ به k بستگی دارد را در نظر می گیریم

$$(Xk\boldsymbol{\beta}+\gamma_0\mathbf{1}_p-Y)'(Xk\boldsymbol{\beta}+\gamma_0\mathbf{1}_p-Y)+\lambda\boldsymbol{\beta}'\boldsymbol{\beta}.$$

با مشتق گرفتن نسبت به β و برابر صفر قرار دادن آن بدست می آوریم :

$$\left[k_{xx}'S_{XX}k_{xx}+\lambda I_p\right]\boldsymbol{\beta}=k_{xx}'S_{XY}$$

بنابراین براورد رگرسیون ریج برای β داده می شود توسط

$$\tilde{\boldsymbol{\beta}}_n(\lambda)=\left[\lambda I_p+(\hat{k}_{xx}'S_{XX}\hat{k}_{xx})^{-1}\right]^{-1}\tilde{\boldsymbol{\beta}}_n$$

که در آن

$$\tilde{\boldsymbol{\beta}}_n=\left(S_{XX}-n\Sigma_{uu}+2\Sigma_{xu}\right)^{-1}S_{XY}.$$

پس فاکتور ریج $R_n(\lambda)$ براوردگر ریج بصورت ذیل نشان داده می شود :

$$R_n(\lambda)=\left[I_p+\lambda C_n^{-1}\right]^{-1},\quad C_n=\hat{k}_{xx}'S_{XX}\hat{k}_{xx}$$

از این رو براوردگر رگرسیون ریج با توجه به تعریف فاکتور ریج β نمایش داده می شود با :

$$\tilde{\boldsymbol{\beta}}_n(\lambda) = R_n(\lambda)\tilde{\boldsymbol{\beta}}_n.$$

فاکتور ریج معرفی شده براوردگر سازگاری است $R_n(\lambda) \xrightarrow{p} R(\lambda)$ وقتی $n \to \infty$ میل می کند پس داریم

$$R(\lambda) = \left[I_p + \lambda C^{-1}\right]^{-1}, \quad C = k'_{xx}\Sigma_{XX}k_{xx}$$
$$\Sigma_{XX} = \Sigma_{xx} + \Sigma_{uu}2\Sigma_{xu}$$

حال اریبی و MSE (میانگین مربعات خطا) بترتیب وقتی $n \to \infty$ میل می کند عبارتند از :

$$b\left(\tilde{\boldsymbol{\beta}}_n(\lambda)\right) = -\lambda C^{-1}(\lambda)\boldsymbol{\beta}, \qquad C^{-1}(\lambda) = (C + \lambda I_p)^{-1}$$
$$\mathrm{MSE}\left(\tilde{\boldsymbol{\beta}}_n(\lambda)\right) = \sigma_{zz}[R(\lambda)]'C^{-1}[R(\lambda)] + \lambda^2 C^{-1}(\lambda)\boldsymbol{\beta}\boldsymbol{\beta}'C^{-1}(\lambda).$$

۳ بحث و نتیجه گیری

در این مقاله سعی شده براورد سازگاری برای پارامتر مدل بدست آوریم وقتی که در مدل خطای اندازه گیری متغیر ابزاری داریم و رگرسورهای مدل دچار همخطی چندگانه اند. اگر از روش کمترین مربعات پارامترها را براورد کنیم براوردهای ضعیف و ناپایداری بدست می آید. با مقدار اریبی که براوردگر ریج می پذیرد در عوض MSE کوچکتری نسبت به براوردگر کمترین مربعات می دهد که یکی از دلایل استفاده از رگرسیون ریج در تجزیه و تحلیل هاست . واضح است که اگر ضریب $\lambda = 0$ درنظر بگیریم براورد ریج تبدیل به براورد کمترین مربعات معمولی می شود.

مراجع

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. Annals of Statistics, 32 (2): 409–499.

Hoerl, A.E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12: 55-67

Fuller, W. A. (1987), MeasurementErrorModels, New York: John Wiley

Fuller, W. A., and Hidiroglou, M. A. (1978), "Regression Estimation After Correcting for Attenuation," Journal of the American Statistical Association,73,99-104

L.J. Gleser, (1992) , The Importance of assessing measurement reliability in multivariate regression, Journal of the American Statistical Association 87 (419) , 696–707.

Shalabh, Gaurav Garg, Neeraj Misra, Use of prior information in the consistent estimation of regression coefficients in a measurement error model, Journal of Multivariate Analysis 100 (2009) 1498–1520.

A.K.Md. Ehsanes Saleh , Shalabh (2014), A ridge regression estimation approach to the measurement error model Journal of Multivariate Analysis 123 , 68–84

Mahajan, A. (2005): "Identification and Estimation of Regression Models with Misclassification," Econometrica, 74(3), 631–665.

Hausman, J., J. Abrevaya, and F. Scott-Morton (1998): "Misclassification of the Dependent Variable in a Disrete-response Setting," Journal of Econometrics, 87, 239–269.