

بررسی اثرات هم خطی در مدل‌های رگرسیونی چندگانه

زهرة طغرای^۱

چکیده

یکی از اهداف رگرسیون چندگانه تعیین اثر هر یک از متغیرهای مستقل با ثابت نگه داشتن سایر متغیرهای مستقل می‌باشد. این هدف در مرحله اول با برآورد ضرایب رگرسیونی در مدل محقق می‌گردد. اما اگر بین متغیرهای مستقل رابطه خطی برقرار باشد جواب یکتایی برای پارامترها قابل دستیابی نیست. در این صورت مشکل هم خطی در مدل رگرسیونی به وجود آمده و محقق در برآورد صحیح پارامترها دچار مشکل می‌گردد. مشکلات ناشی از هم خطی در یک مدل رگرسیونی روی آریبی ناشی از حذف متغیر و نیز واریانس پارامترها می‌باشد. در این مقاله چگونگی این اثرات بررسی و نشان داده شده که هم خطی تحت شرایط خاصی می‌تواند باعث افزایش و یا کاهش دو مورد فوق گردد. علاوه بر این ملاک‌های تشخیص و نیز راه‌های مواجهه با هم خطی بررسی شده‌اند.

واژه های کلیدی: رگرسیون، هم خطی، رگرسیون ستیغی، عامل تورم واریانس، تحلیل مؤلفه‌های اصلی

(برای سادگی در بیان، بدون از دست دادن کلیت، متغیرهای استاندارد شده استفاده می‌گردند).

برآورد بردار β به صورت زیر می‌باشد:

$$\hat{\beta} = (X'X)^{-1}(XY)$$

$$= \begin{bmatrix} n & r_{12} & nr_{1p} \\ nr_{21} & n & nr_{2p} \\ \vdots & \vdots & \vdots \\ nr_{p1} & nr_{p2} & n \end{bmatrix}^{-1} \begin{bmatrix} nr_{y1} \\ nr_{y2} \\ \vdots \\ nr_{yp} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & r_{12} & r_{1p} \\ r_{21} & 1 & r_{2p} \\ \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & 1 \end{bmatrix}^{-1} \begin{bmatrix} ry_1 \\ ry_2 \\ \vdots \\ ry_p \end{bmatrix}$$

$$\Rightarrow \hat{\beta} = R_{xx}^{-1}R_{xy} ; i, j = 1, \dots, p \quad (2)$$

۱- مروری بر رگرسیون چند متغیره و بررسی هم خطی بین متغیرها

یکی از اهداف مهم رگرسیون چندگانه این است که اثر هر یک از متغیرهای مستقل را با ثابت نگهداشتن بقیه متغیرهای مستقل، از هم تفکیک کند که این هدف، در مرحله اول با برآورد ضرایب رگرسیونی در جدول زیر، محقق می‌گردد.

$$Y = X\beta + e \quad (1)$$

Y : برداری شامل n مشاهده از متغیر وابسته.

X : ماتریس $n \times p$ از متغیرهای مستقل.

β : بردار $p \times 1$ شامل پارامترهای جامعه.

e : بردار $n \times 1$ شامل خطاها با ویژگی $e \approx N(0, \delta^2 I)$

^۱ محقق بانک مرکزی جمهوری اسلامی ایران

$$r_{ij} = \text{همبستگی بین } x_i \text{ و } x_j$$

$$r_{yi} = \text{همبستگی بین } x_i \text{ و } y$$

$$n = \text{تعداد مشاهدات}$$

حال اگر بین x ها رابطه خطی برقرار باشد و ارون ماتریس (XX) وجود نداشته و جواب یکتایی برای $\hat{\beta}$ قابل دستیابی نیست. در این صورت مشکل هم خطی در مدل رگرسیون به وجود می آید که محقق در برآورد پارامترها دچار مشکل می گردد.

معادله (۲) نشان می دهد که $\hat{\beta}$ تحت تأثیر همبستگی بین متغیرهای مستقل قرار می گیرد. جانستون^۱ مطرح کرد که اثر حذف j متغیر در معادله (۱)، باعث اریب شدن برآوردهای پارامترهای مربوط به $p-j$ متغیر باقی مانده می شود. این اریبی به صورت $R_{XZ} \beta_Z - I_{XX} R_{XZ}$ می باشد که:

R_{XX} ماتریس همبستگی بین $p-j$ متغیر مستقل باقی مانده.

R_{XZ} همبستگی بین j متغیر حذف شده و $p-j$ متغیر مستقل باقی مانده.

β_Z پارامترهای جامعه مربوط به متغیرهای حذف شده، در مدل رگرسیونی شامل همه متغیرهای مستقل. مشاهده می شود در صورت عدم همبستگی بین متغیرهای مستقل، مقدار اریبی صفر است.

در پنج حالت کلی می توان مدل رگرسیونی و مشکلات مربوط به هم خطی در آن را بررسی کرد:

حالت اول:

زمانی که جامعه داده ها و مدل تئوری مشخص است. در این حالت برآورد رگرسیونی با توجه به کل جامعه و با تمام متغیرهای مناسب پیش بینی می شود. برآوردهای پارامترها همه نااریب بوده و هم خطی هیچگونه مشکلی ایجاد نمی کند.

حالت دوم:

داده های جامعه مشخص نبوده و مدل مناسب و صحیحی برآورد نشده است که در این حالت کل مدل زیر سؤال است.

حالت سوم:

داده های جامعه در دسترس بوده ولی مدل مناسبی بدست نیامده، که در این حالت برآوردهای پارامترهای رگرسیونی اریب هستند.

حالت چهارم:

زمانی که محقق مدل مناسبی با توجه به داده های نمونه تعیین می کند که برآورد پارامترها حتی با وجود هم خطی، نااریب هستند. اما هم خطی می تواند باعث افزایش واریانس برآورد شود. البته در این مقاله نشان داده خواهد شد که هم خطی در این حالت می تواند باعث کاهش واریانس برآوردها نیز بشود.

حالت پنجم:

متداول ترین حالتی که اتفاق می افتد، مشخص نبودن مدل و جامعه است که در این حالت معمولاً داده ها به اندازه کافی، کامل نیستند که همه متغیرهای مربوطه را وارد کنند و مباحث نمونه گیری مطرح است. تحت این شرایط، اریبی ناشی از حذف متغیر و هم تورم (یا کاهش) واریانس برآوردها، به علت هم خطی می تواند مشکل ساز شود.

به طور کلی، موارد ذیل می توانند نشان دهنده وجود هم خطی باشند:

(۱) تغییرات بزرگ در b_k در زمان اضافه و یا

حذف یک یا چند متغیر (یا مشاهده).

(۲) b_k برای متغیر X_k که از نظر تئوری مهم

است، معنی دار نمی باشد.

(۳) علامت b_k با توجه به یافته های قبلی و یا از

نظر تئوری، برخلاف انتظار است.

(۴) همبستگی های بزرگ در ماتریس R_{XX} .

(۵) بزرگ شدن $S(b_k)$ ، انحراف معیار برآورد

پارامتر X_k .

(۶) با وجود معنی دار بودن F مربوط به مدل،

b_k ها معنی دار نیستند.

با توجه به موارد بالا، می توان گفت اگر دو

متغیر مستقل، کاملاً بهم مرتبط باشند، وقتی یکی از

¹John stone

برای ارزیابی اثر هم‌خطی استفاده کنیم. ریشه VIF، عاملی که به وسیله آن خطای معیار و فاصله اطمینان و به علت هم‌خطی چندگانه، افزایش پیدا کردن را بیان می‌کند.

VIF اندازه مناسب برای بررسی وجود هم‌خطی در مورد متغیرهای مستقل است اما اگر متغیرهای مستقل، شامل متغیرهای نشانگر باشند، نمی‌تواند هم‌خطی را ارزیابی کند زیرا همبستگی بین متغیرهای نشانگر به وسیله انتخاب گروه مرجع، متأثر می‌شود. ملاکی که در این حالت مورد استفاده قرار می‌گیرد:

$$GVIF_1 = \frac{\det R_{11} \det R_{22}}{\det R},$$

R_{11} : ماتریس همبستگی X_1 (مجموعه ۱)

R_{22} : ماتریس همبستگی X_2 (مجموعه ۲)

R : ماتریس همبستگی شامل همه متغیرها

مزیت GVIF این است که مستقل از انتخاب گروه مرجع برای متغیر کیفی در مدل رگرسیونی است. با این حال، انتخاب یک گروه بزرگ نتایج با ثبات‌تری بدست می‌آورد. مشابه ریشه گرفتن از VIF مقدار $GVIF^{1/2df}$ ، کاهش در دقت برآوردها را نشان می‌دهد که ناشی از هم‌خطی می‌باشد. (درجه آزادی، تعداد متغیرهای مستقل است).

یکی دیگر از ملاک‌های تشخیص هم‌خطی تولرانس^۲ است. تولرانس برای متغیر X_k برابر است با:

$$(Tol)_k = 1 - R_k^2; \quad k = 1, \dots, p,$$

R_k^2 : R^2 مربوط به مدل رگرسیونی است که در آن مدل X_k متغیر وابسته و متغیرهای مستقل دیگر به عنوان متغیر مستقل هستند. بنابراین می‌توان گفت:

$$(VIF)_k = \frac{1}{(Tol)_k},$$

آنها ثابت نگهداشته می‌شود، ضرایب حداقل مربعات به صورت یکتا تعریف نمی‌شوند و در حالت کلی، ارتباط قوی بین X ها باعث بی‌ثبات شدن ضرایب حداقل مربعات، انحراف معیارهای بزرگ و بی‌دقت شدن برآوردهای β می‌شود. در قسمت بعد به بررسی برخی از ملاک‌های تشخیص هم‌خطی که در نرم‌افزارهای آماری نیز قابل استخراج است می‌پردازیم.

۲- ملاک‌های تشخیص هم‌خطی

با توجه به ماتریس همبستگی متغیرهای مستقل و ضرایب همبستگی دودو می‌توان وجود هم‌خطی را تأیید یا رد کرد. همبستگی دودو بالا، یک هم‌خطی را پیشنهاد می‌کند اما، هم‌خطی چندگانه حتی زمانی ممکن است اتفاق بیفتد که همبستگی‌های دودو به اندازه کافی بالا نباشد. از طرفی دیگر با توجه به R^2 تمام مدل‌های رگرسیونی شامل یک X به عنوان متغیر وابسته و بقیه متغیرهای مستقل، به عنوان متغیرهای مستقل مدل، می‌توان گفت $R^2 \geq b$ وجود یک هم‌خطی را تأیید می‌کند اما ساده‌ترین اندازه برای بررسی وجود هم‌خطی و در مورد تک‌تک متغیرها، ملاک "عامل تورم واریانس" است که در اکثر نرم‌افزارها به صورت VIF^1 نشان داده می‌شود که درون فرمول واریانس برآورد پارامتر است:

$$Var(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_e^2}{(n-1)S_j^2},$$

R_j^2 : R^2 مدل X_j (به عنوان متغیر وابسته) با سایر X ها.

$$S_j^2 = \frac{\sum (X_{ij} - \bar{X}_j)^2}{n-1},$$

جمله $\frac{1}{1 - R_j^2}$ به عنوان VIF نامیده می‌شود. R_j^2 در VIF،

از همبستگی‌های دودو نیست بلکه از همبستگی‌های چندگانه X_j روی X های دیگری می‌باشد زیرا همبستگی‌های دودو، همیشه مسئله هم‌خطی را بیان نمی‌کند. از این‌رو معمولاً «هم‌خطی چندگانه» بیشتر مدنظر است.

از آنجایی که اندازه فاصله اطمینان برای β_j ، متناسب با ریشه VIF است، بهتر است برای راحتی، ریشه VIF را

² Tolerance

¹Variance Inflation Factor

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

$$= \begin{bmatrix} n & nr_{12} & nr_{1p} \\ nr_{p1} & nr_{p2} & n \end{bmatrix}^{-1} \begin{bmatrix} nr_{y1} \\ nr_{y2} \\ nr_{yp} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & r_{12} & r_{1p} \\ r_{11} & 1 & r_{2p} \\ r_{p1} & r_{p2} & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{y1} \\ r_{y2} \\ r_{yp} \end{bmatrix},$$

$$\Rightarrow \hat{\beta} = R_{xx}^{-1}R_{xy}. \quad (2)$$

ماتریس واریانس-کوواریانس $\hat{\beta}$ به صورت زیر است:

$$E(\hat{B} - B)(\hat{B} - B)'$$

$$= \hat{\sigma}^2 (X'X)^{-1} \sum_{\hat{B}} \quad (3)$$

$$= \hat{\sigma}^2 \left(\frac{1}{n} R_{xx}^{-1} \right),$$

$\hat{\sigma}^2$: برآورد σ^2 و به صورت زیر تعریف می‌شود:

$$\hat{\sigma}^2 = S^2 = \text{Var}(Y|X)$$

$$= \left(\frac{n}{n-P} \right) (1 - R'_{xy} R_{xx}^{-1} R_{xy}) \quad (4)$$

$$= \left(\frac{n}{n-P} \right) \frac{|R|}{|R_{xx}|},$$

که $R_{(P+1)(P+1)}$ ماتریس همبستگی بین متغیر وابسته و متغیرهای مستقل می‌باشد.

با استفاده از معادلات ۲-۴، آماره t برای پارامتر p م برابر است با:

$$t_{\hat{\beta}_p} = \frac{\hat{\beta}_p}{\sqrt{\sum_{\hat{\beta}_{pp}}}} \quad (5)$$

که در آن $\sum_{\hat{\beta}_{pp}}$ عنصر p ام قطر برآورد ماتریس واریانس کوواریانس برای برآورد پارامتر p م در معادله ۲ می‌باشد. در حالتی که داده‌ها استاندارد

همانطور که قبلاً هم گفتیم، برای سادگی، از مدل رگرسیون استاندارد شده استفاده می‌کنیم که در آن تمام متغیرها (X ها و Y) به مقادیر Z با میانگین صفر و انحراف معیار یک استاندارد شده و بر $\sqrt{n-1}$ تقسیم می‌شوند. در این مدل، معادلات نرمال $X'Xb = X'Y$ به صورت زیر می‌باشند:

$$R_{xx} b^* = r_{yx}$$

می‌توان نشان داد که $(VIF)_k$ ، k امین عنصر قطر ماتریس $(R_{xx})^{-1}$ است طوری که:

$$\sigma^2(b_k^*) = (\sigma^{*2})(VIF)_k = (\sigma^{*2})(1 - R_k^2)$$

σ^{*2} واریانس مدل استاندارد شده است. بنابراین VIF ، مقدار افزایش واریانس ضرایب رگرسیونی b_k^* مدل استاندارد شده به وسیله هم‌خطی را اندازه می‌گیرد.

به عنوان قاعده کلی $Tol < 0.1$, $VIF > 10$ نشان دهنده، مشکل ساز بودن هم‌خطی می‌باشند. ملاک VIF برای کل مدل به صورت زیر بدست می‌آید:

$$VIF = \frac{\sum_{i=1}^{P-1} VIF_i}{P-1}.$$

ملاک دیگری که برای تشخیص هم‌خطی استفاده می‌شود، ریشه‌های مشخصه از تحلیل مؤلفه‌های اصلی هستند اگر این مقادیر تقریباً به یک اندازه باشند، هم‌خطی مشکل جدی نیست در هم‌خطی‌های کامل، ریشه‌های مشخصه برابر صفر است.

۳- مشکلات ناشی از هم‌خطی در یک مدل

رگرسیونی

همانطور که قبلاً هم دیدیم برآورد پارامترهای β در یک مدل رگرسیونی به صورت $\hat{\beta} = (X'X)^{-1}(X'Y)$ است که آن را به شکل زیر نیز می‌توان نوشت:

r_{12} : همبستگی نمونه‌ای بین X_1 و X_2 است. به خاطر آورید که اریبی ناشی از حذف متغیر به صورت $R_{xx}^{-1} R_{xz} \hat{\beta}_z$ می‌باشد. در نتیجه، می‌توان نشان داد که حذف X_j ضریب رگرسیونی X_i را به اندازه $r_{12} \hat{B}_j$ اریب می‌کند. زمانی که $r_{12} = 0$ ، هیچ‌گونه اریبی ناشی از حذف متغیر وجود ندارد و با یافته‌های جانستون نیز سازگاری دارد. اما در صورت وجود هم‌خطی، اریبی ناشی از حذف متغیر وجود دارد. اخیراً ثابت شده که نمی‌توان گفت تحت شرایط مختلف این اریبی افزایش یا کاهش می‌یابد.

اگر اریبی در برآورد \hat{B}_i را با $bias_i = \hat{B}_j r_{12}$ نشان دهیم مشتق جزئی نسبت به r_{12} ، روند افزایشی یا کاهش‌ی این اریبی را نسبت به تغییرات r_{12} ، بیان می‌کند.

$$\frac{\partial bias_i}{\partial r_{12}} = \frac{r_{y_j} (1 + r_{12}^2) - 2r_{y_i} r_{12}}{(1 - r_{12}^2)^2}$$

افزایش یا کاهش اریبی را در صورت حذف متغیر X_j و درسه حالت زیرمورد بحث قرار می‌دهیم:

الف) برای $r_{y_i} > 0$ ، r_{y_j} ، اریبی \hat{B}_i ، وقتی $r_{12} < 0$ ، شدیداً افزایش می‌یابد و اگر $r_{12} > 0$ باشد، اریبی شدیداً افزایش (کاهش) می‌یابد اگر

$$\frac{r_{y_j}}{r_{y_i}} > \left(\frac{2r_{12}}{1 + r_{12}^2} \right)$$

ب) اگر $r_{y_i} < 0$ ، r_{y_j} ، اریبی \hat{B}_i ، وقتی $r_{12} < 0$ ، شدیداً کاهش می‌یابد و وقتی $r_{12} > 0$ ،

این اریبی افزایش (کاهش) می‌یابد اگر

$$\frac{r_{y_j}}{r_{y_i}} < \left(\frac{2r_{12}}{1 + r_{12}^2} \right)$$

می‌شوند، $SST = n$ ، از طرفی $S^2 = \frac{e'e}{n-P}$ و $SSE = s^2(n-P)$ است. بنابراین داریم:

$$SSE = s^2(n-P) = \frac{n}{n-P} (1 - R'_{xy} R_{xx}^{-1} R_{xy})(n-P) \quad (6)$$

$$= n \frac{|R|}{|R_{xx}|}$$

بنابراین R^2 به صورت زیر تعریف می‌شود:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = 1 - \frac{|R|}{|R_{xx}|} \quad (7)$$

آماره F نیز به صورت زیر تعریف می‌شود:

$$F_{P-1, n-P} = \left(\frac{n-P}{P-1} \right) \left(\frac{R^2}{1-R^2} \right) = \frac{(|R_{xx}| - |R|)(n-P)}{|R| \times (P-1)}$$

مشاهده می‌شود تمام نتایج در تحلیل یک مدل رگرسیونی P به نوعی به همبستگی بین متغیرهای مستقل، مربوط می‌شود. اما در حالت کلی، هم‌خطی، دو تأثیر مهم در یک مدل رگرسیونی می‌تواند داشته باشد که در زیر به شرح آن پرداخته می‌شود [۱].

۴- تأثیر هم‌خطی روی اریبی ناشی از حذف متغیر

مدل رگرسیونی استاندارد شده و با دو متغیر مستقل را به صورت زیر در نظر می‌گیریم:

$$Y = \beta_1 X_1 + \beta_2 X_2 + e; e \approx N(0, \sigma^2)$$

از آنجایی که متغیرها استاندارد شده‌اند، مقدار ثابتی در مدل وجود ندارد. در این مدل برآورد پارامترها به صورت زیر می‌باشد:

$$\hat{B} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix} = \begin{bmatrix} \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2} \\ \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2} \end{bmatrix} \quad (10)$$

r_{yj} : همبستگی نمونه‌ای بین Y و متغیر مستقل X_j است $(j = 1, 2)$.

به عبارتی هم خطی، تحت شرایط خاصی، واریانس برآورد پارامترها را کاهش می دهد. برای مثال زمانی که $r_{yi} < 0$ و $r_{yj} > 0$ و $D > 0.5$ و $r_{yy} > 0$ و $r_{yi} < 0$ (در جهت منفی)، واریانس برآورد پارامتر کاهش می یابد.

تمام ملاک های تشخیص، بیان می کنند وقتی همبستگی منفی است، واریانس برآورد پارامتر، افزایش می یابد در حالیکه در شرایطی نیز کاهش می یابد [۵].

۶- راه های مواجهه با مشکل هم خطی

با وجود اینکه مشاهده شد، تحت شرایط خاصی، هم خطی می تواند مشکل ساز نباشد، اما بهرحال در حالت کلی باید به نحوی، آن را از بین برد. راهکارهای مختلفی در این مورد پیشنهاد شده که به صورت زیر می باشند: [۲] و [۴].

۱. مادامی که در محدوده مقادیر متغیر مستقل، هدف از پردازش مدل، پیش بینی های \hat{Y}_n یا $\hat{Y}_{h(new)}$ باشد، هم خطی مشکلی به وجود نمی آورد.
۲. زمانی که متغیرهای هم خط به طور توأم معنی دار نباشند، می توانند راحت از مدل حذف شوند.
۳. اگر در یک مدل چند جمله ای، هم خطی داشته باشیم، بهتر است X_i ها را به $(X_i - \bar{X})$ تبدیل کنیم.
۴. در بعضی از موارد با داشتن داده های اضافی، الگوی هم خطی از بین می رود. البته این پیشنهاد، در داده های آزمایشی بهتر از مطالعات مشاهده ای، توجیه می شود.
۵. در بعضی موارد، می توان b_k را از مجموعه دیگری از داده ها برآورد کرده و اثر X_k متناظرش در مدل با استفاده از تبدیل $Y_i' = Y_i - b_k X_k$ از بین می رود.
۶. بر مبنای افزایش واریانس ضرایب برآورد شده در اثر هم خطی، روش رگرسیون استیجی^۱ اریبی کوچکی را برای کاهش $S(b_k)$ معرفی می کند.

ج) اگر $r_{yi} < 0$ ، $r_{yj} > 0$ ، اریبی \hat{B}_i ، وقتی $r_{12} > 0$ ، شدیداً افزایش می یابد و وقتی $r_{12} < 0$ این اریبی افزایش (کاهش) می یابد اگر

$$\frac{r_{yj}}{r_{yi}} < \left(\frac{2r_{12}}{1+r_{12}^2} \right)$$

بنابراین یک اریبی نامتقارن ناشی از حذف متغیر و با توجه به ساختار ضرایب همبستگی، وجود دارد - برای مثال زمانی که $r_{yi} = r_{yj} = r_y > 0$ در $r_{12} = 0.5$ شیب $\left| \frac{\partial bias_i}{\partial r_{12}} \right|$ برابر $r_y = 0.67$ و در $r_{12} = -0.5$ برابر $r_y = 4$ می باشد - که در این حالت ملاک تشخیص VIF، دچار مشکل می شود. طوری که اریبی ناشی از حذف متغیر در صورت همبستگی منفی، بدتر از حالت همبستگی مثبت است اما VIF در هر دو حالت یک مقدار دارد.

بنابراین مشاهده می شود وجود هم خطی بسته به علامت این همبستگی خطی و شرایط مختلف، تأثیرات متفاوتی روی اریبی ناشی از حذف متغیر دارد که ملاک های تشخیص قادر به تفکیک آن ها نیستند.

ثابت شده اریبی ناشی از حذف متغیر می تواند حتی زمانی که VIF مقادیر پایینی دارند نیز بزرگ باشد [۳].

۵- اثر هم خطی روی واریانس برآورد پارامتر

برای بیان اثر هم خطی روی واریانس برآورد پارامتر، با استفاده از معادلات ۴-۳ و (۱۰) داریم:

$$Var(\hat{B}_1) = Var(\hat{B}_1) = \frac{(1-r_{y_1}^2 - r_{y_2}^2 - r_{12}^2 + 2r_{y_1}r_{y_2}r_{12})}{(n-2)(1-r_{12}^2)^2}$$

وجود r_{12} در معادله بالا، نشان می دهد که همبستگی های مثبت و منفی تأثیرات مختلفی روی واریانس مورد نظر دارند. مانند حالت قبل می توان نشان داد:

واریانس برآورد پارامترهای X_i و X_j هر دو با r_{12} افزایش (کاهش) می یابند اگر:

$$r_{y_i} r_{y_j} > \left(\frac{(1-r_{12}^2 - 2D)}{(1-r_{12}^2)} \right)$$

که $D \in [0, 1]$ و دترمینان ماتریس همبستگی Y و X_1 و X_2 است.

¹ Ridge Regression

توجه داشته باشید که هر مؤلفه، یک جمع وزنی از متغیرهای اصلی (X) است که در آن a_{ij} ، وزن یا ضریب مربوط به متغیر i و مؤلفه j باشد. شکل ماتریس رابطه بالا به صورت زیر است:

$$W = Z_x A,$$

هدف از این تحلیل، ساختن اولین مؤلفه‌ای است که تا حد امکان واریانس را تبیین کند. بعد از یافتن اولین مؤلفه، به بررسی و محاسبه مؤلفه‌هایی ادامه می‌دهیم که با یکدیگر متعامد باشند به این معنی که:

$$\sum_{i=1}^P a_{ij}^2 = 1 \quad ; \quad j = 1, \dots, P,$$

$$\sum_{i=1}^P a_{ij} a_{ik} = 0 \quad ; \quad j \neq k, j = 1, \dots, P,$$

$$\sum_{ij} a_{ij} a_{ik} = 0 \quad ; \quad k = 1, \dots, P,$$

برقراری شرایط فوق، این اطمینان را می‌دهد که موقعیت‌های نسبی داده‌ها، بدون تغییر باقی می‌ماند. جمع واریانس این مؤلفه‌های جدید برابر است با جمع واریانس متغیرهای اصلی

$$\sum_{j=1}^P Var(W_j) = \sum Var(X_i),$$

از آنجایی که اولین مؤلفه، بزرگترین میزان واریانس کل را تبیین می‌کند، این مؤلفه، بهترین خلاصه یک بعدی از داده‌هاست.

۸. روش دیگر مقابله با مشکل هم‌خطی، «تحلیل عاملی» است. این روش نیز مانند PCA، چندین متغیر را با تعداد کمتری از متغیرهای جدید خلاصه می‌کند با این تفاوت که تحلیل عاملی مدلی ارائه می‌دهد که طبق آن یک متغیر پنهان می‌سازد در حالیکه تحلیل مؤلفه‌های اصلی، فقط در کاهش تعداد متغیرهای مستقل موفق است. تحلیل مؤلفه‌های اصلی برای تعیین تعداد فاکتورها در تحلیل عاملی، قابل استفاده است.

در تحلیل عاملی، واریانس به دو قسمت تجزیه می‌شود: قسمت اشتراکی (واریانس X_i) که به وسیله

هدف در این روش، بدست آوردن برآوردگری می‌باشد تا با احتمال بالایی، نزدیک مقدار واقعی پارامتر باشد. اندازه «احتمال نزدیک بودن به مقدار واقعی پارامتر» اثرات آریبی و تغییر نمونه‌گیری را ترکیب کرده و «میانگین مربع خطا» نامیده می‌شود.

$$E(b^R - \beta)^2 = \sigma^2(b^R) + (E\{b^R\} - \beta)^2,$$

b^R : برآوردگر ستیغی است که به صورت زیر بدست می‌آید:

$$b^R(\theta) = (X'X + \theta I)^{-1} X'Y,$$

$\theta \geq 0$ یک مقدار ثابت است. معمولاً مقادیر $0 \leq \theta \leq 1$ مناسب هستند. این برآوردگر، ناریب نمی‌باشد. این مدل رگرسیونی، مجموعه‌ای از ضرایب را می‌یابد که ثبات بیشتری داشته و میانگین مربع خطای کوچکی داشته باشد.

برای بدست آوردن b^R ، به صورت ضمنی یک مقدار اولیه تعیین می‌کنیم. این کار معمولاً با رسم نمودار b^R در مقابل θ انجام می‌شود. این نمودار «اثر ستیغی» نامیده می‌شود. θ بی انتخاب می‌شود که:

• ضرایب b^R با ثباتی بدست آیند.

• $(VIF)_k$ به طور معنی‌داری کوچک شوند.

بعد از بدست آوردن b_k ، می‌توان ضرایب معمولی و غیراستاندارد را به صورت زیر بدست آورد:

$$b_k = \left(\frac{S_y}{S_k} \right) b_k^R,$$

$$b_0 = y - b_1 x_1 - \dots - b_p x_p,$$

S_k و S_y انحراف معیارهای X_k و Y می‌باشند.

۷. یکی دیگر از روش‌های برطرف کردن اثر هم‌خطی «تحلیل مؤلفه‌های اصلی» یا PCA است. PCA، P متغیر همبسته را با تعداد کوچک‌تری از متغیرهای ناهمبسته جانشین می‌کند. این کار براساس ماکزیمم کردن میزان کل واریانس (حجم واریانس‌های P متغیر) براساس متغیرهای جدید انجام می‌شود. مؤلفه‌های اصلی، ترکیبات خطی از متغیرهای اصلی به شکل استاندارد شده هستند:

$$W_1 = a_{11}z_1 + a_{12}z_2 + \dots + a_{1p}z_p,$$

$$W_2 = a_{21}z_1 + a_{22}z_2 + \dots + a_{2p}z_p,$$

$$W_p = a_{p1}z_1 + a_{p2}z_2 + \dots + a_{pp}z_p,$$

این روش زمانی قابل استفاده است که تعداد متغیرها خیلی زیاد نباشد اگر P ، تعداد متغیرهای مستقل در رگرسیون باشد، γ^P مدل ممکن وجود دارد. هدف یافتن بهترین زیرمجموعه ممکن با بهترین برازش و کمترین متغیرهای مستقل می‌باشد. یکی از ملاک‌هایی که در این روش استفاده می‌شود C_p است که به صورت زیر محاسبه می‌شود:

$$C_p = \frac{\sum E_i^{\gamma}}{S_E^{\gamma}} + \gamma P - n$$

$$= (k + 1 - P)(f_p - 1) + P$$

S_E^{γ} : یک مدل کامل شامل k متغیر مستقل است.
 F_p : آزمون F برای این فرض بکار می‌رود که متغیرهای مستقلی از زیرمجموعه حذف شوند که شیب صفر داشته باشند. اگر فرض درست باشد:

$$E_{(CP)} \cong P, \quad E_{(JP)} \cong 1$$

ماکزیمم کردن C_p ، جمع مربعات باقی‌مانده را می‌نیمم و R^{γ} را ماکزیمم می‌کند. در یک مدل خوب C_p به P نزدیک است. از نموداری که این دو را در مقابل یکدیگر رسم می‌کند، می‌توان استفاده کرد.

۷- خلاصه و نتیجه گیری

همانطور که در روش هدونیک اشاره شد، هم‌خطی شدید بین متغیرهای مستقل مدل رگرسیونی، تحلیل مدل رگرسیونی را با مشکل مواجه می‌نماید و در بخش دوم راه‌هایی برای کاهش اثر هم‌خطی ارائه شد که با توجه به متغیرهای مستقل و هدف از برازش مدل می‌توان ساده‌ترین روش کاهش اثر هم‌خطی را انتخاب نمود ولی در حالت کلی با افزایش حجم نمونه می‌توان تا حدی بر این مشکل غلبه نمود.

بنابراین با کمینه نمودن هم‌خطی بین متغیرهای مستقل می‌توان به مدلی با ضرایب دقیق‌تری دست یافت که این امر به رسیدن به اهداف برازش مدل کمک بزرگی می‌نماید.

فاکتورهای معمولی و مشترک تبیین می‌شود) و σ_i^{γ} (واریانس باقی‌مانده).

در اینجا هدف این است که تا حد امکان، تغییرات موجود در متغیرهای مستقل (X) با یک مجموعه کوچک‌تری از متغیرهای جدید توضیح داده شود. برازش یک مدل تحلیل عاملی به وسیله سه معیار زیر تعیین می‌شود:

- ۱) درصد واریانس توضیح داده شده به وسیله فاکتورها.
- ۲) مقایسه ماتریس همبستگی دوباره تولید شده با ماتریس همبستگی متعامد. می‌توان یک آزمون نسبت درستنمایی برای این فرض بسازیم که ماتریس همبستگی مناسبی وجود دارد که عدم رد این فرض، یک برازش خوب را نشان می‌دهد.
- ۳) خطای معیار بارهای عاملی (ضرایب تصویر شده X ها روی فاکتورها).

۹. روش دیگر، استفاده از تکنیک انتخاب متغیرهاست. این روش، تعداد متغیرهای مستقل را تا اندازه‌ای کاهش می‌دهد تا مجموعه‌ای با کمترین هم‌خطی به وجود آید. یکی از این تکنیک‌ها، روش گام به گام^۱ است. روش گام به گام خود به دو شکل این مجموعه را به وجود می‌آورد: الف) روش گام به گام پیشرو: متغیرهای مستقل را یکی یکی به مدل اضافه می‌کند. در هر مرحله متغیری که بزرگترین افزایش را در R^{γ} به وجود می‌آورد، انتخاب می‌شود و این کار زمانی متوقف می‌شود که میزان افزایش از ملاک از پیش تعیین شده، کمتر باشد.

ب) روش گام به گام پسرو: به صورت مشابه، با این تفاوت که با یک مدل کامل شروع شده و یکی یکی متغیرها را حذف می‌کند.

ج) روش پسرو/پیشرو که ترکیب دو روش است. ایرادی که به تکنیک گام به گام وارد است این است که یک تغییر (اصلاح) کوچک در داده‌ها، یا یک نمونه جدید، باعث عکس شدن نتایج می‌شود. به عبارتی مدل‌های رگرسیونی پیشرو و پسرو ممکن است نتایج متفاوتی بدست آورند.
 ۱۰. تکنیک تعیین زیرمجموعه:

¹Stepwise

منابع

- [1]. Mela, C. F. and Kopalle, P. K. (2002). The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlation. Applied Economics, Vol, 34, No. 6, 20, pp. 667-677.
- [2]. Anderson, B. (1990). Regression: advanced method lecture 9. Detecting and handling collinearity. McMASTER University .
- [3]. Hendricks, J., Belzer, B., Grotendis, M. and Lammers, J. (1999). Collinearity involving ordered and unordered categoriocal variables.
- [4]. Multiple regression /STAT 5616 spring (2001). Chapter 9, Collinearity.
- [5]. Kuh. B. and Welsch (1980). Multicollinearity.