رگرسیون لجستیک حداقل مربعات جزیی

مرضیه شاهمندی، سروش علیمرادی، علی رجالی
گروه آمار، دانشگاه صنعتی اصفهان

رگرسیون لجستیک چندگانه با متغیرهای توضیحی به هم وابسته در دامنه وسیعی از علوم از جمله علوم اجتماعی، اقتصادی، مهندسی، کشاورزی و پزشکی کاربرد دارد. در مدل رگرسیون لجستیک چندگانه اگر متغیرهای توضیح دهنده وابسته باشند، آن‌گاه مدل ناپایدار شده و برآورد پارامترهای مدل، بسیار نادقیق می‌شود و حتی ممکن است تفسیر رابطه بین متغیر پاسخ و هر متغیر توضیح دهنده با استفاده از نسبت‌های بخت نادرست باشد. از طرفی افزایش تعداد متغیرهای توضیح دهنده موجب وجود همبستگی بین متغیرها می‌گردد و لذا محققین برای رفع این مشکلات روش‌هایی را پیشنهاد می‌دهند. یکی از این روش‌ها تعمیم روش حداقل مربعات جزیی به رگرسیون لجستیک چندگانه است. که با کمک روش‌های بوت استرپ، معنی داری ضرایب این نوع از رگرسیون بررسی می‌شوند. در این مقاله ابتدا رگرسیون خطی تعمیم یافته جزیی معرفی می‌شود. سپس با به‌کارگیری مجموعه داده واقعی مربوط به یک نوع ورق فولاد تولیدی، رگرسیون لجستیک حداقل مربعات جزیی به‌کار گرفته می‌شود و با رگرسیون لجستیک معمولی مقایسه می‌گردد و نتیجه می‌شود که مدل رگرسیون لجستیک حداقل مربعات جزیی نسبت به مدل رگرسیون لجستیک معمولی پایدارتر است و از جهت پیشگویی بهتر عمل می‌کند و تمامی متغیرهای توضیح دهنده تأثیر گذار در آن حضور دارند.

واژه‌های کلیدی: رگرسیون لجستیک چندگانه، هم‌خطی چندگانه، رگرسیون حداقل مربعات جزیی، بوت استرپ.

۱ مقدمه

رگرسیون حداقل مربعات جزیی [۱] (PLS) به مطالعه رابطه بین یک متغیر پاسخ عددی و یک مجموعه k تایی از متغیرهای توضیح دهنده در شرایطی که رگرسیون چندگانه ناپایدار و نشدنی است، می پردازد. از مواردی که امکان استفاده رگرسیون چندگانه وجود ندارد:

[1] Partial Least Squares

۱ ـ وجود هم‌خطی چندگانه بین متغیرهای توضیح دهنده
۲ ـ کوچک بودن مشاهدات در مقایسه با تعداد متغیرهای توضیح دهنده
۳ ـ وجود داده گمشده در بین مشاهدات

امکان برخورد با چنین مشکلاتی در مواقع به‌کارگیری رگرسیون لجستیک یا به‌طور کلی‌تر مدل‌های خطی تعمیم یافته نیز وجود دارد. رگرسیون PLS توسط ولد (۱۹۸۴) معرفی شد. اما اول بار در علم شیمی توسط گلادی و کولاسکی (۱۹۸۶) به‌کار رفت و معروف شد. کاربرد رگرسیون حداقل مربعات جزیی در علوم مربوط به ارزیابی حواس در کارهای مارتنز و نایس (۱۹۸۹) نیز دیده شده است. این نوع از رگرسیون توسط فرانک و فریدمن (۱۹۹۳) و تننهاس (۱۹۹۸) در چهارچوب آمار تفسیر شد. باستین و همکاران (۲۰۰۵) رگرسیون تعمیم یافته جزیی و الگوریتم مربوط به آن را مطرح کردند. همچنین روشی جهت محاسبه مولفه‌های PLS و تعیین تعداد این مولفه‌ها برای باقی ماندن در مدل ارائه دادند. هانگ و همکارش (۲۰۰۹)، روش PLS غیر خطی را برای مواقعی که رابطه همبستگی بین متغیرهای توضیح دهنده و متغیرهای پاسخ غیر خطی است پیشنهاد کردند. در این مقاله رگرسیون تعمیم یافته جزیی و الگوریتم محاسبه مولفه‌های PLS مرتبط با آن تشریح می‌شود. در ادامه تنها به رگرسیون لجستیک PLS از این کلاس رگرسیونی، بسنده می‌شود. سپس با به‌کارگیری یک مجموعه داده واقعی مربوط به یک نوع ورق فولاد تولیدی، رگرسیون لجستیک معمولی و رگرسیون لجستیک PLS مقایسه می‌شوند. همچنین روش‌های بوت استرپ جهت اعتبار سنجی ضرایب رگرسیون لجستیک PLS به کار می‌روند. و در انتها نتیجه گیری بیان می‌شود.

۲ رگرسیون خطی تعمیم یافته جزیی

فرض کنید $X_۱, X_۲, ..., X_p$ یک مجموعه از متغیرهای پیوسته که بدون خطا مشاهده می‌شوند، باشد. n مشاهده از هر کدام از متغیرها را در نظر گرفته و در ماتریس $\mathcal{X} = (x_{ij})_{n\times p} = \{\boldsymbol{x}_۱, \boldsymbol{x}_۲, \dots, \boldsymbol{x}_p\}$ قرار می‌دهیم. بردار $\mathcal{Y} = (y_۱, y_۲, ..., y_n)'$ یک نمونه تصادفی از متغیر پاسخ Y که مرتبط با مشاهدات ماتریس $\mathcal{X}$ است، را فرض کنید. با این محدودیت که مولفه‌های PLS، $\boldsymbol{t}_h = \sum_{j=۱}^{p} w^*_{hj}\boldsymbol{x}_j$ ها متعامد هستند مدل رگرسیون خطی تعمیم یافته جزیی [۲] $(PLS-GLR)$ $\mathcal{Y}$ روی $\boldsymbol{x}_۱, \boldsymbol{x}_۲, \dots, \boldsymbol{x}_p$ با m مولفه به‌صورت زیر

[2] PLS Generalised Linear Regression

تعریف می‌شود (باستین و همکاران، ۲۰۰۵)

$$g(\Theta) \quad = \quad \sum_{h=1}^{m} c_h (\sum_{j=1}^{p} w^*_{hj} \boldsymbol{x}_j)$$

که پارامتر Θ می‌تواند میانگین یک متغیر پاسخ پیوسته Y یا بردار احتمال مقادیر قابل تعریف یک متغیر گسسته Y باشد. تابع ارتباط g با توجه به توزیع احتمالی Y و نکویی برازش مدل به داده‌ها انتخاب می‌شود.

۲-۱ الگوریتم $(PLS-GLR)$

این الگوریتم شامل ۴ قدم است.

قدم ۱ : محاسبه m مولفه PLS، $\boldsymbol{t}_h$ برای $h=1,2,\ldots,m$.

قدم ۲ : رگرسیون خطی تعمیم یافته $\mathcal{Y}$ روی m مولفه PLS باقی مانده.

قدم ۳ : بیان $(PLS-GLR)$ بر حسب متغیرهای توضیح دهنده اصلی.

قدم ۴ : اعتبار سنجی متقابل بوت استرپ ضرایب در مدل نهایی $(PLS-GLR)$.

قدم اول و آخر به‌طور کامل شرح داده می‌شود و در ادامه ۲ قدم دیگر در قالب یک مثال روشن می‌شود.

فرض کنید $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_p\}$ ماتریس متغیرهای توضیح دهنده استاندارد شده است. هدف پیدا کردن m مولفه متعامد PLS، $\boldsymbol{t}_h$ است.

محاسبه اولین مولفه PLS، $\boldsymbol{t}_1$:

قدم ۱ : محاسبه ضریب رگرسیونی a_{1j}، متغیر توضیح دهنده $\boldsymbol{x}_j$ با استفاده از رگرسیون خطی تعمیم یافته $\mathcal{Y}$ روی $\boldsymbol{x}_j$ برای هر متغیر $\boldsymbol{x}_j$، $j=1,2,\ldots,p$.

قدم ۲ : نرمال کردن بردار ستونی a_1 شامل اعضای a_{1j}، $w_1=\frac{a_1}{\|a_1\|}$.

قدم ۳ : محاسبه مولفه $\boldsymbol{t}_1$ وقتی $\boldsymbol{t}_1=\frac{\mathcal{X}w_1}{w_1'w_1}$.

محاسبه دومین مولفه PLS، $\boldsymbol{t}_2$:

قدم ۱ : محاسبه ضریب رگرسیونی a_{2j}، متغیر توضیح دهنده $\boldsymbol{x}_j$ با استفاده از رگرسیون خطی تعمیم یافته $\mathcal{Y}$ روی متغیرهای $\boldsymbol{t}_1$ و $\boldsymbol{x}_j$ برای هر متغیر $\boldsymbol{x}_j$، $j=1,2,\ldots,p$.

قدم ۲ : نرمال کردن بردار ستونی a_2، با اعضای a_{2j}، $w_2=\frac{a_2}{\|a_2\|}$.

قدم ۳ : محاسبه ماتریس باقی مانده $\mathcal{X}_1$ از رگرسیون خطی $\mathcal{X}$ روی مولفه $\boldsymbol{t}_1$ وقتی $\boldsymbol{x}_{h-1,j}$، j امین ستون ماتریس $\mathcal{X}_{h-1}$ است.

قدم ۴ : محاسبه مولفه $\boldsymbol{t}_2$ وقتی $\boldsymbol{t}_2=\frac{\mathcal{X}_1 w_2}{w_2'w_2}$.

قدم ۵ : بیان مولفه $\boldsymbol{t}_2$ بر حسب متغیرهای توضیح دهنده اصلی، $\boldsymbol{t}_2 = \mathcal{X}w_2^*$.

$\vdots$

محاسبه h امین مولفه PLS، $\boldsymbol{t}_h$:

قدم ۱ : محاسبه ضریب رگرسیونی a_{hj} متغیر توضیح دهنده $\boldsymbol{x}_j$ با استفاده از رگرسیون خطی تعمیم یافته $\mathcal{Y}$ روی متغیرهای $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{h-1}$ و $\boldsymbol{x}_j$ برای هر متغیر $\boldsymbol{x}_j$، $j = 1, 2, \ldots, p$.

قدم ۲ : نرمال کردن بردار ستونی a_h، با اعضای a_{hj}، $w_h = \frac{a_h}{\|a_h\|}$.

قدم ۳ : محاسبه ماتریس باقی مانده $\mathcal{X}_{h-1}$ از رگرسیون خطی $\mathcal{X}$ روی مولفه‌های $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{h-1}$ وقتی $\boldsymbol{x}_{h-1,j}$، j امین ستون ماتریس $\mathcal{X}_{h-1}$ است.

قدم ۴ : محاسبه مولفه $\boldsymbol{t}_h$ وقتی $\boldsymbol{t}_h = \frac{\mathcal{X}_{h-1}w_h}{w_h' w_h}$.

قدم ۵ : بیان مولفه $\boldsymbol{t}_h$ بر حسب متغیرهای توضیح دهنده اصلی، $\boldsymbol{t}_h = \mathcal{X}w_h^*$.

۲-۲ ساختن معادله رگرسیون لجستیک PLS

در انتها، یک مدل رگرسیون لجستیک $\mathcal{Y}$ روی مولفه‌های $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_h$ مطرح می‌شود. معادله رگرسیون لجستیک PLS با بیان این معادله به عنوان یک تابعی از متغیرهای توضیح دهنده اصلی به‌دست می‌آید. اگر Y متغیر پاسخ دو تایی و π احتمال پیشامد $(Y = 1)$ در نظر گرفته شود، آن‌گاه داریم :

$$
\begin{aligned}
ln(\widehat{\frac{\Pi}{1-\Pi}}) &= c_1\boldsymbol{t}_1 + c_2\boldsymbol{t}_2, \ldots, c_h\boldsymbol{t}_h \\
&= c_1\mathcal{X}w_1^* + c_2\mathcal{X}w_2^* + \ldots + c_h\mathcal{X}w_h^* \\
&= \mathcal{X}b
\end{aligned}
$$

که $b = c_1w_1^* + c_2w_2^* + \ldots + c_hw_h^*$ و $\Pi = (\pi_1, \pi_2, \ldots, \pi_n)'$.

۲-۳ نکته‌ها

۱ ـ در محاسبه مولفه PLS $(\boldsymbol{t}_h)$، به‌جای ضریب رگرسیونی a_{hj} که معنی دار نیست، صفر قرار می‌دهیم یا به‌عبارتی در محاسبه مولفه PLS تنها متغیرهای معنی دار شرکت دارند.

۲ ـ تنها h مولفه در مدل باقی می‌ماند هرگاه برای محاسبه $(h+1)$ امین مولفه $(\boldsymbol{t}_{h+1})$،

هیچ یک از ضرایب $a_{h+1,j}$ تفاوت معنی داری از صفر نداشته باشند.
۳ ـ الگوریتم بیان شده در صورت وجود داده گمشده نیز کاربرد دارد.

مثال ۱ . هدف مورد نظر بررسی تأثیر خواص مکانیکی و شیمیایی بر یک عیب خاص در ورق‌های فولاد تولید شده توسط یک کارخانه تولید ورق فولاد است. این مجموعه داده شامل ۵۰ مشاهده از اطلاعات تولید در خصوص متغیر پاسخ (Y) و متغیرهای توضیح دهنده است.
Y : یک متغیر دودویی است که $y = ۱$ اگر ورق فولاد تولیدی سالم و $y = ۰$ به تولید معیوب اطلاق می‌شود و همچنین $x_۱$: استحکام تسلیم ورق فولاد (بر حسب نیوتن بر میلی متر مربع)، $x_۲$: استحکام کشش نهایی ورق فولاد (بر حسب نیوتن بر میلی متر مربع)، $x_۳$: فلز سیلیکن (بر حسب درصد)، $x_۴$: فلز آلومینیم (بر حسب درصد)، $x_۵$: فلز منیزیم (بر حسب درصد) هستند.

برازش رگرسیون لجستیک معمولی

در ادامه تمام محاسبات روی پنج متغیر توضیح دهنده استاندارد شده انجام می‌گیرد.
نتایج برازش رگرسیون لجستیک معمولی متغیر پاسخ روی پنج متغیر توضیح دهنده استاندارد شده در جدول (۱) و پاسخ‌های مشاهده شده و پیشگویی شده حاصل از این برازش در جدول (۲) گزارش شده است.

پارامتر	درجه آزادی	برآورد پارامتر	انحراف استاندارد	کای اسکور والد	$p-$ مقدار
ثابت	۱	۲/۴۱۶	۰/۶۶	۱۳/۴۰۱۰	۰/۰۰۰۳
$x_۱$	۱	−۰/۴۶۹۱	۰/۷۳۲۹	۰/۴۰۹۶	۰/۵۲۲۲
$x_۲$	۱	−۱/۱۹۰۸	۰/۵۶۳۹	۴/۴۵۹۸	۰/۰۳۴۷
$x_۳$	۱	−۰/۶۰۷۷	۰/۵۶۷۹	۱/۱۴۵۱	۰/۲۸۴۶
$x_۴$	۱	۰/۵۴۵۴	۰/۵۸۰۲	۰/۸۸۳۷	۰/۳۴۷۲
$x_۵$	۱	۱/۱۰۲۲	۰/۶۴۹۲	۲/۸۸۲۷	۰/۰۸۹۵

جدول ۱. آنالیز حداکثر درستنمایی برآوردها.

تعدا پاسخ مشاهده شده	پیشگویی شده ۰	۱	مجموع
۰	۲	۶	۸
۱	۲	۴۰	۴۲
مجموع	۴	۴۶	۵۰

جدول ۲. پیشگویی پاسخ بر اساس مدل رگرسیون لجستیک معمولی.

مشاهده می‌شود که تعداد پاسخ به اشتباه کلاس بندی شده ۸ تا است که ۱۶% از پاسخ‌های مشاهده شده است.
در جدول (۳)، همبستگی پیرسون بین متغیرهای توضیح دهنده استاندارد شده گزارش شده است که علامت ستاره همبستگی‌های معنی دار در سطح اعتبار ۰/۰۵ را نشان می‌دهد.

	x_1	x_2	x_3	x_4	x_5
x_1	۱				
x_2	−۰/۰۶۶۶	۱			
x_3	−۰/۳۵۸۳*	۰/۲۹۵۶*	۱		
x_4	۰/۳۸۳۹*	−۰/۱۰۴۸	۰/۴۲۹۱*	۱	
x_5	۰/۴۳۶۵*	۰/۰۰۹۳	−۰/۰۲۷۳	۰/۱۰۳۸	۱

جدول ۳. همبستگی‌های پیرسون بین متغیرهای استاندارد شده.

برازش رگرسیون لجستیک PLS

در داده‌های ورق فولاد، وجود هم‌خطی‌های معنی دار بین تعدادی از متغیرهای توضیح دهنده منجر به مشکلات زیر می‌شود.
۱ ـ با توجه به جدول (۱) متغیرهای تأثیر گذار x_4، x_3 و x_1 در سطح اعتبار ۰/۱، نا معنی دار شده اند.
۲ ـ پاسخ‌های به اشتباه کلاس بندی شده، درصد بالایی را شامل می‌شوند.
در ادامه برای برازش رگرسیون لجستیک PLS نیاز به محاسبه مولفه‌های PLS است. با توجه به روش پیشنهادی این مولفه‌ها محاسبه می‌شوند.
برای به‌دست آوردن مولفه اول،t_1، رگرسیون‌های لجستیک جداگانه

پاسخ روی هر کدام از متغیرهای توضیح دهنده استاندارد شده انجام می‌شود. ضرایب a_{1j} به‌ترتیب (۰/۰۴۸۴) ۰/۹۷۸۸− و (۰/۰۸۸۴) ۰/۶۵۷۵− و (۰/۰۵۹) ۰/۷ و (۰/۰۱۱) ۱/۱۶۹۹ و (۰/۰۹۹) ۰/۶۹۷۸− هستند. مقدار بیان شده در پرانتز pـ مقدار[3] است. ضرایب ذکر شده همگی در سطح اعتبار ۰/۱ معنی دار هستند. بعد از نرمال کردن ضرایب، اولین مولفه PLS به‌فرم زیر محاسبه می‌شود:

$$\begin{aligned} \boldsymbol{t}_1 &= \frac{-0.6978\boldsymbol{x}_1 + 1.1699\boldsymbol{x}_2 + 0.7\boldsymbol{x}_3 - 0.6575\boldsymbol{x}_4 - 0.9788\boldsymbol{x}_5}{\sqrt{0.6978^2 + 1.1699^2 + 0.7^2 + 0.6575^2 + 0.9788^2}} \\ &= -0.361\boldsymbol{x}_1 + 0.6053\boldsymbol{x}_2 + 0.3622\boldsymbol{x}_3 - 0.3402\boldsymbol{x}_4 - 0.5064\boldsymbol{x}_5 \end{aligned}$$

برای محاسبه دومین مولفه PLS، t_2، ابتدا متغیرهای توضیح دهنده‌ای که سهم معنی داری در محاسبه t_2 دارند مشخص می‌شوند. برای این کار رگرسیون‌های لجستیک پاسخ روی $\boldsymbol{t}_1$ و هر متغیر توضیح دهنده استاندارد شده انجام می شود.

مقادیر pـ مقدار مربوط به ضرایب این متغیرهای توضیح دهنده استاندارد شده به‌ترتیب ۰/۷۸۳۷ و۰/۶۸۸۴ و ۰/۷۹۰۹ و ۰/۳۸۹۹ و ۰/۱۷۰۹ است که هیچ کدام از سطح اعتبار ۰/۱ کمتر نیستند. بنابراین دومین مولفه PLS معنی دار نیست یا به‌عبارتی دیگر مدل با یک مولفه PLS باقی می‌ماند. با بیان این مولفه بر حسب متغیرهای توضیح دهنده، مدل به فرم زیر حاصل می‌شود:

$$Pr(\mathcal{Y}=1) = \frac{e^{2.17+0.4283\boldsymbol{x}_1-0.718\boldsymbol{x}_2-0.4297\boldsymbol{x}_3+0.4036\boldsymbol{x}_4+0.6\boldsymbol{x}_5}}{1+e^{2.17+0.4283\boldsymbol{x}_1-0.718\boldsymbol{x}_2-0.4297\boldsymbol{x}_3+0.4036\boldsymbol{x}_4+0.6\boldsymbol{x}_5}}$$

مشاهده می‌شود که تمامی متغیرهای توضیح دهنده تأثیر گذار در این مدل حضور دارند و همچنین تعداد پاسخ‌های به اشتباه کلاس بندی شده بر اساس این مدل ۷ تا است یعنی یکی کمتر از تعداد پاسخ‌های به اشتباه کلاس بندی شده حاصل از مدل رگرسیون لجستیک معمولی. با توجه به کاهش سال‌های به اشتباه کلاس بندی شده، برآورد پارامترهای مدل لجستیک از روش PLS مناسب‌تر از رگرسیون لجستیک معمولی است.

[3] P-Value

۴-۲ اعتبار سنجی برای رگرسیون لجستیک PLS

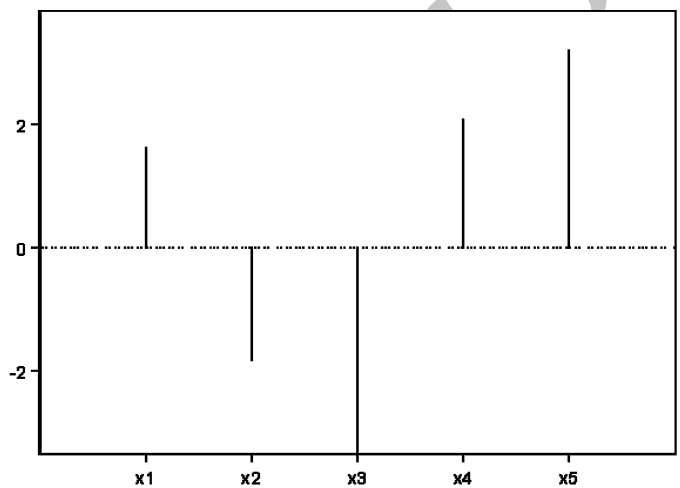
بوت استرپ تصادفی و متعادل

به‌منظور به‌دست آوردن توزیع تجربی برای پارامترهای رگرسیون لجستیک PLS، یک نوع ساده از بوت استرپ به نام بوت استرپ تصادفی در نظر گرفته می‌شود. ایده این نوع از بوت استرپ اصلاح تابع توزیع تجربی است.
(ایفرون، ۱۹۸۳) این بوت استرپ را برای حالت پاسخ گسسته Y دوتایی $(K = ۲)$ پیشنهاد داد. فرض کنید $\hat{F}$ توزیع تجربی (y_i, x_i) باشد. تعمیم و توسعه بوت استرپ تصادفی به رگرسیون لجستیک PLS (تعمیم به حالتی که $K > ۲$ است) به فرم زیر است.
فرض کنید $\pi_i(y_i = k|x_i)$، احتمال مشاهده $y_i = k \ \ , k = ۱, ۲, \ldots, K$ به شرط بردار سطری از متغیرهای توضیح دهنده، x_i باشد. این احتمالات از دو راه متفاوت قابل محاسبه هستند:
۱ ـ برآوردهای احتمال شرطی از رگرسیون لجستیک معمولی یا رگرسیون لجستیک PLS.
۲ ـ استفاده از توزیع تجربی Y.
فرض کنید B تعداد نمونه‌های بوت استرپ باشد. بوت استرپ تصادفی قدم‌های زیر را شامل می‌شود:
قدم ۱ : برای هر $b = ۱, ۲, \ldots, B$، فرض کنید $\mathcal{X}^{(b)}$، b امین نمونه بوت استرپی از n متغیر تصادفی iid تولید شده از $\hat{F}^{RAND}$ باشد. برای تعیین یک مقدار مناسب $y_i^{(b)} = k \ \ , k = ۱, ۲, \ldots, K$ برای هر بردار $x_i \ \ , i = ۱, ۲, \ldots, n$، یک مقدار شبیه سازی شده از $\pi_i(y_i = k|x_i)$ را در نظر بگیرید.
قدم ۲ : برای هر نمونه بوت استرپی $(\mathcal{Y}^{(b)}, \mathcal{X}^{(b)})$، برآورد ضرایب رگرسیون لجستیک PLS، $\hat{\beta}^{(b)}$، به‌دست می‌آید.
قدم ۳ : برای هر متغیر توضیح دهنده $x_j \ \ , j = ۱, ۲, \ldots, p$ تقریب مونت کارلوی توزیع بوت استرپ $\hat{\beta}_j$، $\Theta^{(B)}_{\hat{F}^{RAND}, \hat{\beta}_j}$، به‌دست می‌آید.
در انتها، صدک‌های $\Theta^{(B)}_{\hat{F}^{RAND}, \hat{\beta}_j}$ جهت به‌دست آوردن فواصل اطمینان برای β_j، به‌کار گرفته می‌شود.
ذکر این نکته اهمیت دارد که کارایی محاسبات بوت استرپ، توسط بوت استرپ متعادل یا به‌عبارتی همان بوت استرپ تقدم و تأخر (ایفرون و تیب شیرانی، ۱۹۹۳)، که یک اصلاحیه و تعدیل ساده از ساختار نمونه گیری بوت استرپ معمولی است، افزایش می‌یابد.
جهت انجام بوت استرپ متعادل قدم‌های زیر انجام می‌شود:
۱ ـ B کپی از $\mathcal{X}$ را در یک ماتریس Υ با nB سطر قرار دهید.
۲ ـ یک تقدم و تأخر تصادفی بین سطرهای Υ ایجاد کنید.

۳ ـ n سطر اول ماتریس Υ، اولین نمونه بوت استرپی، n سطر دوم ماتریس Υ، دومین نمونه بوت استرپی و به همین ترتیب، n سطر آخر ماتریس Υ، b امین نمونه بوت استرپی در نظر گرفته می‌شود.

بوت استرپ متعادل مشابه یک طرح بلوکی ناقص متعادل عمل می‌کند. این نوع طرح‌ها با توجه به فرم ماتریس اطلاع متناظرشان، طرح‌های متعادل کارا هستند و با توجه به ملاک‌های بهینگی، جز طرح‌های بهینه به‌شمار می آیند.

روش اعتبار سنجی ناپارامتری پیشنهاد شده برای رگرسیون لجستیک PLS شامل یک مولفه با استفاده از داده‌های ورق فولاد با $B = ۱۰۰۰$ به‌کار گرفته شده است و نتیجه در شکل (۱) گزارش شده است. با توجه به فواصل اطمینان نشان داده شده در شکل (۱) و عدم وجود صفر در این فواصل، همه متغیرها در مدل معنی دار هستند.



شکل (۱) : فواصل اطمینان ۰٫۹۵ بوت استرپ تصادفی متعادل برای داده‌های ورق فولاد

۳ نتیجه گیری

مشاهده می‌شود که در صورت وجود هم‌خطی چندگانه بین متغیرهای توضیح دهنده، به‌کارگیری رگرسیون لجستیک PLS نسبت به رگرسیون لجستیک معمولی برآوردهای

مناسب‌تری برای پارامترها به‌دست می‌دهد. همچنین این مدل همه متغیرهای توضیح دهنده تأثیر گذار را شامل می‌شود و از جهت پیشگویی متغیر پاسخ، بهتر عمل می‌کند.

مراجع

[1] Bastien, P. , Esposito Vinzi, V. and Tenenhaus, M. (2005). PLS generalized linear regression, *Comput. Statist. Data Anal.*, **48**, No. 1, 17-46.

[2] Efron, B. (1983). Estimating the error rate of a prediction rule: some improvments on cross-validation, *J. Amer. Statist. Assoc.*, **78**, 316-331.

[3] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to The Bootstrap*, Chapman and Hall, New York.

[4] Frank, I.E., and Fridman, J.H. (1993). A statistical view of chemometrics regression tools, *Technometrics*, **35**, 109-148.

[5] Geladi, P., and Kowlaski, B. (1986). Partial least square regression: A tutorial, *Analytica. Chemica.*, **35**, 1-17.

[6] Martens, H., and Naes, T. (1989). *Multivariate calibration*, London: Wiley.

[7] Tenenhaus, M. (1998). *La Re̒gression PLS*, Paris: Technip.

[8] Wold, H. (1984). PLS regression, *Encyclopadia of Statistical Sciences*, Academic Press, New York, **6**, 581-591.

[9] Zhang, Yi., and Zhang, Y., (2009). Complex process monitoring using modified partial least squares method of independent component regression, *Chemom. Intell. Labora Syst.*, **98**, 143-148.