گزیده‌مطالب آماری
سال ۲۰، شماره‌ی ۲، پاییز و زمستان ۱۳۸۸، صص ۲۸۹-۳۰۴

رگرسیون کم‌ترین توان‌های دوم

امیر شاهینی* و مجید سرمد

دانشگاه فردوسی مشهد

چکیده: در تحقیقات مختلف به‌طور معمول با مسائلی سر و کار داریم که با استفاده از مجموعه‌ای از متغیرهای توضیحی به پیش‌بینی رفتار یك متغیر وابسته می‌پردازیم. یکی از روش‌های آماری که کاربرد وسیعی در این‌گونه مسائل دارد رگرسیون چندگانه می‌باشد. اما هنگامی که بین متغیرهای توضیحی رابطه‌ی خطی وجود داشته باشد مسئله‌ی هم‌خطی چندگانه رخ می‌دهد و در نتیجه رگرسیون کم‌ترین توان‌های دوم معمولی به ایجاد براوردهای نااستواری از ضرایب رگرسیونی می‌انجامد. رگرسیون کم‌ترین توان‌های دوم جزئی یک روش چند متغیره است که در هنگام بروز هم‌خطی بین متغیرهای توضیحی مورد استفاده قرار می‌گیرد. در این مقاله به معرفی این روش می‌پردازیم.

۱- مقدمه

رگرسیون کم‌ترین توان‌های دوم جزئی (Partial Least Squares Regression) که به اختصار آن را رگرسیون PLS نیز می‌نامند، یک روش آماری است که اخیراً در زمینه‌های مختلف به خصوص در مسائل شیمی‌سنجی محبوبیت فراوانی پیدا کرده است ([۳] و [۷]). این روش در سال ۱۹۶۰ میلادی توسط یک اقتصاددان سوئیسی به نام هرمن ولد در زمینه اقتصادسنجی معرفی و به کار برده شد [۹]. هنگامی که تعداد متغیرهای توضیحی بسیار زیاد است و بین متغیرها هم‌خطی شدید وجود دارد رگرسیون PLS در ساختن مدل‌هایی برای پیش‌بینی متغیر پاسخ بسیار مفید است. شایان ذکر است که PLS یک روش رگرسیونی اریب است که بیش‌تر در پیش‌بینی متغیر پاسخ استفاده می‌شود و معمولاً کاربردی در تفسیر ارتباط بین متغیر پاسخ و متغیرهای توضیحی ندارد

واژگان کلیدی: اعتبارسنجی متقابل، پیش‌بینی، هم‌خطی چندگانه، مؤلفه، PLS.
دریافت: ۱۳۸۸/۶/۱۷، پذیرش: ۱۳۸۸/۱۰/۲
* نویسنده‌ی عهده‌دار مکاتبات

[۵].همچنین هنگامی که بیش از یک متغیر پاسخ در مسئله وجود دارد نیز می‌توان از این روش استفاده نمود. این روش بر اساس الگوریتم‌های نسبتاً پیچیده‌ای تعریف شده که این موضوع فهم آن را کمی مشکل می‌کند. در این مقاله ابتدا تفسیر ساده‌ای از این الگوریتم در حالت یک متغیره بیان می‌شود. سپس با استفاده از شبیه‌سازی این روش با بعضی از روش‌های رگرسیونی دیگر مورد مقایسه قرار می‌گیرد.

فرض کنیم یک نمونه به حجم n برای بررسی ارتباط خطی بین متغیر پاسخ Y و متغیرهای توضیحی X_1، X_2، ...، X_m داریم. iامین مشاهده، $i = 1,2,\ldots,n$، به صورت $\left(x_{1}(i), x_{2}(i), \ldots, x_{m}(i), y(i)\right)$ و بردارهای مقادیر مشاهده‌شده‌ی Y و X_j، $j = 1,2,\ldots,m$، به صورت $\mathbf{y} = \left\{y(1), y(2), \ldots, y(n)\right\}'$ و $\mathbf{x}_j = \left\{x_j(1), x_j(2), \ldots, x_j(n)\right\}'$ نمایش داده می‌شوند. همچنین میانگین‌های نمونه‌ای آن‌ها نیز به صورت $\bar{y} = \sum_i y(i)/n$ و $\bar{x}_j = \sum_i x_j(i)/n$ نشان داده می‌شوند.

۲- PLS یک‌متغیره

در این روش برای تشکیل یک رابطه بین متغیر Y و متغیرهای توضیحی X_1، X_2، ...، X_m، متغیرهای توضیحی جدیدی ساخته می‌شود که آن‌ها را فاکتور (factor)، متغیر پنهان (latent variable) یا مؤلفه (component) می‌نامند. هر یک از این متغیرهای جدید یک ترکیب خطی از متغیرهای اولیه X_1، X_2، ... ، X_m می‌باشد. سپس از روش‌های رگرسیونی استاندارد برای تعیین معادلاتی که این مؤلفه‌ها را به متغیر Y ارتباط دهند استفاده می‌شود. در عین حال این روش با به‌کار بردن مؤلفه‌هایی با قدرت پیش‌بینی بالا که تعداد آن‌ها کم‌تر از تعداد متغیرهای اولیه است بُعد مسئله را نیز کاهش می‌دهد. این کاهش بُعد با استفاده از روشی به نام "اعتبارسنجی متقابل" (Cross Validation) صورت می‌گیرد. معادله‌ی رگرسیون حاصل به‌صورت زیر می‌باشد:

$$\hat{Y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_p T_p \qquad (p < m) \qquad (1)$$

که در آن هر مؤلفه‌ی T_k ترکیب خطی از X_j ها است و همچنین همبستگی نمونه‌ای بین هر جفت از این مؤلفه‌ها صفر می‌باشد. در واقع PLS سعی دارد مدلی بسازد که در آن هم‌خطی شدیدی که بین متغیرهای اولیه وجود داشت از بین برود و در عین حال با براورد تعداد پارامترهای کم‌تر بُعد مسئله را نیز کاهش دهد. با این مقدمه یکی از الگوریتم‌های مربوط به این روش را که در سال ۱۹۸۲ توسط ولد [۹] پیشنهاد شد شرح می‌دهیم. شایان ذکر است که الگوریتم‌های مختلفی توسط آماردانان دیگر پیشنهاد شده است که آن‌ها نیز به نتایجی مشابه الگوریتم ولد منجر می‌شوند ([۲]، [۸] و [۶]) الگوریتم ولد به‌صورت زیر است.

گام ۱: $\mathbf{T} \leftarrow \circ$

گام‌های ۲ تا ۵ زیر را m مرتبه تکرار کن.

گام ۲: $b_k \leftarrow \frac{\mathbf{X}_k'.\mathbf{Y}_k}{\mathbf{X}_k'.\mathbf{X}_k}$

گام ۳: $w_k \leftarrow \mathbf{X}_k'.\mathbf{X}_k$

گام ۴: $\hat{\mathbf{Y}}_k \leftarrow w_k b_k \mathbf{X}_k$

گام ۵: $\mathbf{T} \leftarrow \hat{\mathbf{Y}}_k + \mathbf{T}$

گام ۶: $\mathbf{Y} \leftarrow \mathbf{Y} - \frac{\mathbf{T}'.\mathbf{Y}}{\mathbf{T}'.\mathbf{T}}.\mathbf{T}$

گام ۷: $\mathbf{X} \leftarrow \mathbf{X} - \frac{\mathbf{T}'.\mathbf{X}}{\mathbf{T}'.\mathbf{T}}.\mathbf{T}$

گام ۸: اگر $\mathbf{X}'.\mathbf{X} = \circ$ به الگوریتم پایان بده.

الگوریتم بالا مربوط به ساختن مؤلفه‌ها در PLS است. قبل از شروع الگوریتم متغیرهای توضیحی و پاسخ استاندارد می‌شوند. در این صورت اعضای خارج قطر اصلی ماتریس($\mathbf{X}'.\mathbf{X}$) نشان‌دهنده‌ی همبستگی بین متغیرهای توضیحی خواهند بود. در ابتدا متغیر پاسخ روی هر یک از متغیرهای توضیحی به‌طور جداگانه رگرسیون می‌شود (گام ۲) و چون متغیرها استاندارد هستند معادلات پیش‌بینی حاصل به‌صورت زیر می‌باشند:

$$\hat{\boldsymbol{Y}}_k = b_k \boldsymbol{X}_k \qquad k = ۱, ۲, \ldots, m \qquad (۲)$$

که با توجه به روش کم‌ترین توان‌های دوم معمولی داریم $b_k = \frac{\mathbf{x}'_k.\mathbf{y}}{\mathbf{x}'_k.\mathbf{x}_k}$

هرکدام از m معادله‌ی قبل یک براورد برای پاسخ Y به دست می‌آورند. اما در این‌جا از ارتباط‌هایی که بین $\mathbf{X}_k$ ها وجود داشت چشم‌پوشی شده است. حال برای رفع این مشکل با تلفیق کردن این براوردها می‌توانیم میانگین وزنی براوردهای (۲) را در نظر بگیریم (گام های ۳ تا ۵). در این صورت اولین مؤلفه به‌صورت زیر به دست خواهد آمد:

$$\mathbf{T}_1 = \sum_{k=1}^{m} w_k b_k \mathbf{X}_k \qquad (۳)$$

با توجه به حالات مختلفی که می‌توان برای تعیین وزن‌های w_k در نظر گرفت رابطه‌ی (۳) نیز به‌صورت‌های مختلفی ممکن است نوشته شود. در پایان این بخش دو راه برای تعیین w_kها ارائه خواهد شد.

از آن‌جا که $\mathbf{T}_1$ یک میانگین وزنی برای پیشگوکننده‌های $\mathbf{Y}$ (یعنی $\hat{\mathbf{Y}}_k$ ها) است باید خودش نیز یک پیشگوکننده‌ی مفید برای $\mathbf{Y}$ باشد. اما $\mathbf{T}_1$ قادر نیست همه تغییرات $\mathbf{Y}$ را توضیح دهد. آن بخش از تغییرات $\mathbf{Y}$ که توسط $\mathbf{T}_1$ توضیح داده نمی‌شود با استفاده از مانده‌های حاصل از رگرسیون $\mathbf{Y}$ روی $\mathbf{T}_1$ قابل برآورد است (گام ۶). از طرف دیگر متغیرهای توضیحی $\mathbf{X}_k$ به‌طور بالقوه شامل اطلاعات مفیدتری نسبت به $\mathbf{T}_1$ در پیش‌بینی $\mathbf{Y}$ هستند. به این ترتیب آن بخش از اطلاعات موجود در متغیر $\mathbf{X}_k$ که در $\mathbf{T}_1$ وجود ندارد نیز به وسیله‌ی مانده‌های حاصل از رگرسیون $\mathbf{X}_k$ روی $\mathbf{T}_1$ برآورد خواهد شد (گام ۷).

مانده‌های حاصل از رگرسیون $\mathbf{X}_k$ روی $\mathbf{T}_1$ به‌عنوان متغیر توضیحی جدید و مانده‌های حاصل از رگرسیون $\mathbf{Y}$ روی $\mathbf{T}_1$ به‌عنوان متغیر پاسخ جدید در مرحله‌ی بعد برای ساختن مؤلفه‌ی $\mathbf{T}_2$ به کار می‌رود و گام‌های ۱ تا ۵ تکرار می‌شود. این روند برای ساختن مؤلفه‌های بعدی نیز به‌همین صورت انجام می‌شود و هر مؤلفه از مانده‌های حاصل از رگرسیون روی مؤلفه‌ی قبلی خود به دست می‌آید.

حال فرض کنیم $\mathbf{Y}_i$ و $\mathbf{X}_{ik}$ $(k = 1,2,\ldots,m)$ مانده‌های رگرسیونی حاصل از مرحله‌ی $(i-1)$ و $\mathbf{T}_i$ نیز مؤلفه‌ای باشد که در مرحله‌ی iام توسط متغیرهای $\mathbf{Y}_i$ و

$\mathbf{X}_{ik}$ $(k = ۱,۲,\ldots,m)$ ساخته شده است و مقادیر نمونه‌ای مربوط به $\mathbf{T}_i$ ، $\mathbf{Y}_i$ و $\mathbf{X}_{ik}$ نیز به ترتیب به‌صورت $\mathbf{t}_i$ ، $\mathbf{y}_i$ و $\mathbf{x}_{ik}$ باشند.

برای به دست آوردن مؤلفه‌ی $\mathbf{T}_{i+۱}$ نخست باید $\mathbf{Y}_{i+۱}$ و $\mathbf{X}_{(i+۱)k}$ $(k = ۱,۲,\ldots,m)$ تعیین شوند. برای این منظور $\mathbf{X}_{ik}$ $(k = ۱,۲,\ldots,m)$ روی $\mathbf{T}_i$ رگرسیون می‌شود و $\mathbf{X}_{(i+۱)k}$ به‌صورت زیر تعریف می‌شود:

$$\mathbf{X}_{(i+۱)k} = \mathbf{X}_{ik} - \frac{\mathbf{t}_i' . \mathbf{x}_{ik}}{\mathbf{t}_i' . \mathbf{t}_i} . \mathbf{T}_i \qquad k = ۱,۲,\ldots,m \tag{۴}$$

به‌طور مشابه، $\mathbf{Y}_{(i+۱)}$ نیز به‌صورت $\mathbf{Y}_{i+۱} = \mathbf{Y}_i - \frac{\mathbf{t}_i' . \mathbf{y}_i}{\mathbf{t}_i' . \mathbf{t}_i} . \mathbf{T}_i$ تعریف می‌شود.

"تغییرات باقیمانده" در $\mathbf{Y}$ برابر $\mathbf{Y}_{i+۱}$ و "اطلاعات باقیمانده" در $\mathbf{X}_k$ برابر $\mathbf{X}_{(i+۱)k}$ می‌باشد. حال کافیست $\mathbf{Y}_{i+۱}$ روی $\mathbf{X}_{(i+۱)k}$ رگرسیون شود (گام ۲). اگر ضرایب رگرسیونی را با $b_{(i+۱)k}$ نشان دهیم، این ضرایب به‌صورت زیر محاسبه می‌شوند:

$$b_{(i+۱)k} = \frac{\mathbf{x}'_{(i+۱)k} . \mathbf{y}_{(i+۱)}}{\mathbf{x}'_{(i+۱)k} . \mathbf{x}_{(i+۱)k}} \qquad k = ۱,۲,\ldots,m \tag{۵}$$

پس از تعیین وزن‌های $w_{(i+۱)k}$ (گام ۳) می‌توان مؤلفه $\mathbf{T}_{i+۱}$ را به‌صورت زیر به دست آورد:

$$\mathbf{T}_{(i+۱)} = \sum_{k=۱}^{m} w_{(i+۱)k} b_{(i+۱)k} \mathbf{X}_{(i+۱)k} \tag{۶}$$

اما این الگوریتم تا چه زمانی ادامه می‌یابد یا به عبارت دیگر شرط پایان این الگوریتم چیست؟

در پاسخ به این سئوال می‌توان گفت: تا زمانی که ماتریس $\mathbf{X}$ یک ماتریس پوچ (null matrix) شود این الگوریتم ادامه می‌یابد که این شرط به صورت $\mathbf{X}'.\mathbf{X} = ۰$ در گام ۸ آمده است.

پس از این‌که مؤلفه‌ها تعیین شدند با استفاده از مدل رگرسیونی (۱) با متغیر Y ارتباط داده می‌شوند و ضرایب رگرسیونی β نیز به روش کم‌ترین توان‌های دوم معمولی برآورد می‌شوند.

همان‌طور که گفتیم یکی از مزایای PLS ناهمبسته بودن هر جفت از مؤلفه‌ها است زیرا

الف- مانده‌های رگرسیونی با متغیر توضیحی ناهمبسته هستند. مثلاً $\mathbf{X}_{(i+1)k}$ (به‌ازای هر k) با $\mathbf{T}_i$ ناهمبسته است.

ب- هر یک از مؤلفه‌های $\mathbf{T}_p$ ، ...، $\mathbf{T}_{i+1}$ یک ترکیب خطی از $\mathbf{X}_{(i+1)k}$ ها می‌باشد. بنا بر این با توجه به (الف) هر یک از $\mathbf{T}_p$ ، ...، $\mathbf{T}_{i+1}$ با $\mathbf{T}_i$ ناهمبسته‌اند.

اما برای تعیین تعداد مؤلفه‌هایی که باید در مدل (۱) باشند (p) شکلی از روش "اعتبارسنجی متقابل" به کار برده می‌شود که در بخش بعد به توضیح آن خواهیم پرداخت. پس از این‌که مؤلفه‌ها انتخاب شدند مدل رگرسیون (۱) با استفاده از روابط (۴) و (۶) بر اساس متغیرهای اولیه‌ی X بیان می‌شود. در پایان این بخش دو روش برای تعیین وزن‌های w_{ik} مشخص می‌کنیم.

روش اوّل: $w_{ik} = \mathbf{x}'_{ik}.\mathbf{x}_{ik}$، بنابراین $w_{ik} b_{ik} = \mathbf{x}'_{ik}.\mathbf{y}_i$ و چون $\mathbf{x}'_{ik}.\mathbf{y}_i \propto \hat{\text{cov}}(\mathbf{X}_{ik}, \mathbf{Y}_i)$ با استفاده از رابطه‌ی (۶) می‌توان نوشت

$$\mathbf{T}_i = \sum_{k=1}^{m} (w_{ik} b_{ik}) \mathbf{X}_{ik} \propto \sum_{k=1}^{m} \hat{cov}(\mathbf{X}_{ik}, \mathbf{Y}_i) \mathbf{X}_{ik} \qquad (۷)$$

بنا بر رابطه‌ی (۷)، مقدار کوواریانس بین متغیر پاسخ و متغیرهای توضیحی در هر مرحله در ساختن مؤلفه‌های PLS مؤثر هستند. به‌عبارت دیگر، آن متغیرهایی که کوواریانس بیش‌تری با متغیر پاسخ دارند سهم بیش‌تری نیز در پیش‌بینی پاسخ خواهند داشت. بنا بر این می‌توان گفت: PLS در ساختن مؤلفه‌ها هم به ارتباط‌های بین متغیرهای توضیحی و هم به ارتباط‌های بین متغیرهای توضیحی با متغیر پاسخ توجه دارد.

از آن‌جا که $\text{var}(b_{ik}) \propto \{1/(\mathbf{x}'_{ik}.\mathbf{x}_{ik})\}$ پس می‌توان گفت انگیزه‌ی این روش وزن‌دهی، متناسب بودن w_{ik}ها با معکوس واریانس ضرایب رگرسیونی b_{ik} می‌باشد. زیرا در این صورت ضرایبی که دقت بیش‌تری (واریانس کم‌تری) در برآورد پاسخ دارند وزن بیش‌تری نیز در ساختن مؤلفه‌ها خواهند داشت.

روش دوم: $w_{ik} = \frac{1}{m}$، بنا بر این همه‌ی پیشگوکننده‌های پاسخ وزن برابری در ساختن مؤلفه‌ها خواهند داشت.

۳- اعتبارسنجی متقابل (Cross Validation)

این روش برای انتخاب مدلی با بیش‌ترین توانایی پیش‌بینی در بین مدل‌های آماری مورد استفاده قرار می‌گیرد. در واقع برای ارزیابی مدل، توانایی آن را در پیش‌بینی داده‌هایی خارج از محدوده‌ی داده‌هایی که در برازش مدل به کار رفته‌اند مورد قضاوت قرار می‌دهیم. در ادامه این روش را شرح می‌دهیم.

در این روش ابتدا نمونه به k بخش تقسیم می‌شود که حجم این بخش‌ها در صورت امکان تقریباً برابر است. در بین این k بخش یک بخش کنار گذاشته می‌شود که آن را مجموعه‌ی آزمون (test set) می‌نامیم و از اجتماع $(k-1)$ بخش دیگر که مجموعه‌ی آموزشی (training set) نامیده می‌شوند برای "براورد" پارامترهای مدل استفاده می‌شود. سپس با استفاده از این مدل داده‌های مجموعه‌ی آزمون را "پیش‌بینی" کرده و میزان دقت پیش‌بینی این مدل را به‌وسیله‌ی یک تابع زیان مناسب محاسبه می‌کنیم. حال یکی دیگر از k بخش را به‌عنوان مجموعه‌ی آزمون در نظر گرفته و تمام اعمال قبل را تکرار می‌کنیم. بنا بر این هر داده در نمونه‌ی اصلی یک مرتبه پیش‌بینی خواهد شد. توجه کنیم که مجموع k تابع زیان به‌عنوان اندازه‌ای برای دقت پیش‌بینی مدل در نظر گرفته می‌شود. همچنین وقتی مدلی انتخاب شد باید پارامترهای آن را با استفاده از نمونه‌ی اصلی براورد کنیم. معمولاً مقادیر k بین ۲ و ۳ در نظر گرفته می‌شود. هر چند بهتر است مقدار k برابر با حجم نمونه اصلی باشد به طوری‌که هر بخش فقط شامل یک داده باشد. در این صورت براوردهای پارامتر که برای ارزیابی دقت مدل محاسبه می‌شوند تقریباً مشابه براوردهایی هستند که با استفاده از نمونه‌ی اصلی به دست می‌آیند.

مثال ۱- یک نمونه‌ی ۱۰ تایی (x, y) را در نظر می‌گیریم و روش قبل را اعمال می‌کنیم و فرض می‌کنیم که مدل رگرسیونی خطی ساده را در هر بار به داده‌ها برازش می‌دهیم. با توجه به مطالب قبل هر بار ۹ نمونه را به‌عنوان مجموعه‌ی آموزشی در نظر گرفته و با استفاده از مدل به دست آمده مقدار نمونه‌ی حذف‌شده را پیشگویی می‌کنیم.

در جدول ۱ مقادیر براوردها، پیش‌بینی و خطای حاصل از پیش‌بینی در هر مرحله به دست آمده است. مجموع توان‌های دوم خطای پیش‌بینی برای هر مدل را به‌عنوان تابع زیان به‌صورت زیر در نظر می‌گیریم.

جدول ۱- محاسبه‌ی مقدار زیان پس از حذف نمونه‌ی iام برای مدل خطی

i	$x_{(i)}$	$y_{(i)}$	$\hat{\beta}_{(i)0}$	$\hat{\beta}_{(i)1}$	$\hat{y}_{(i)}$	$\{y(i)-\hat{y}(i)\}^2$
۱	۲۲	۹۶	۳۴/۷	۲/۵۶۴	۹۱/۱	۲۴/۳
۲	۲۳/۴	۸۸	۴۷/۸	۲/۱۹	۹۹	۱۲۱/۱
۳	۲۴/۹	۱۰۵	۳۴/۹	۲/۵۵	۹۸/۴	۴۳/۲
۴	۲۸/۵	۱۱۱	۳۸/۳	۲/۴۵۳	۱۰۸/۲	۷/۸
۵	۲۹/۸	۱۰۷	۴۰/۱	۲/۴۱۹	۱۱۲/۲	۲۷/۳
۶	۳۱/۶	۱۱۳	۳۹/۲	۲/۴۴۳	۱۱۶/۴	۱۱/۷
۷	۳۴/۲	۱۳۲	۴۱/۶	۲/۳۱۸	۱۲۰/۹	۱۲۳/۵
۸	۳۶/۴	۱۲۲	۳۶/۱	۲/۵۵۶	۱۲۹/۲	۵۱/۲
۹	۳۷/۷	۱۳۵	۴۲/۱	۲/۳۲	۱۲۹/۶	۲۹/۳
۱۰	۳۹	۱۳۱	۳۶/۲	۲/۵۴۴	۱۳۵/۴	۱۹/۷

$$\text{loss} = \sum_{i=1}^{10} \{y_{(i)} - \hat{y}_{(i)}\}^2$$

در این صورت زیان کل برای این مدل برابر است با

$$۲۴/۳ + ۱۲۱/۱ + ... + ۱۹/۷ = ۴۵۹/۱$$

حال اگر از مدل درجه‌ی دوم $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$ برای این داده‌ها استفاده کنیم و روش قبل را به کار ببریم آن‌گاه زیان کل برای مدل درجه‌ی دوم برابر ۵۶۹/۲ به دست می‌آید که بزرگ‌تر از زیان مدل خطی ساده است. بنا بر این روش اعتبارسنجی متقابل بین این دو مدل، مدل خطی را ترجیح می‌دهد. همچنین توجه کنیم که اگر با همه‌ی داده‌های نمونه مدل را برازش داده و مجموع توان‌های دوم مانده را به‌عنوان معیار تابع زیان در نظر می‌گرفتیم آن‌گاه مدل درجه‌ی دوم انتخاب می‌شد.

اما در روش PLS برای به دست آوردن تعداد مؤلفه‌های موجود در مدل، روشی مشابه روش قبل به کار می‌بریم که چگونگی آن به شرح زیر است.

ابتدا نمونه به k بخش تقسیم می‌شود و یک بخش به‌عنوان مجموعه‌ی آزمون حذف شده و از $(k-1)$ بخش دیگر برای ساختن مؤلفه‌ها با استفاده از الگوریتم PLS استفاده می‌شود. متغیر Y را روی مؤلفه‌های به دست آمده رگرسیون کرده و پارامترهای مدل را به روش کم‌ترین توان‌های دوم براورد می‌کنیم. سپس مقادیر مجموعه آزمون را توسط این مدل پیش‌بینی کرده و خطای حاصل از پیش‌بینی را با استفاده از یک تابع زیان مناسب محاسبه می‌کنیم. این عمل را k مرتبه و هر بار برای یکی از k بخش به‌عنوان مجموعه‌ی آزمون تکرار می‌کنیم و مجموع خطاهای پیش‌بینی در این k مرتبه را به‌عنوان خطای کل مدل در نظر می‌گیریم. حال مؤلفه‌ها را به مدل اضافه می‌کنیم تا این‌که مؤلفه‌ای این خطای کل را افزایش دهد. به این ترتیب تعداد مؤلفه‌هایی که باید در مدل بمانند مشخص خواهد شد.

۴- شبیه‌سازی

در این بخش نحوه‌ی عملکرد PLS و دیگر روش‌های رگرسیونی مورد مقایسه قرار می‌گیرد. در واقع قصد ما تشخیص وضعیت‌هایی است که PLS عملکرد خوبی نسبت به روش‌های دیگر دارد. شایان ذکر است که این شبیه‌سازی بر اساس شبیه‌سازی انجام‌شده در مرجع [۴] صورت گرفته و در پایان، نتایج حاصل از این شبیه‌سازی با نتایج حاصل از شبیه‌سازی مرجع [۴] مورد مقایسه قرار خواهد گرفت. همچنین تمام محاسبات با استفاده از نرم‌افزار R انجام شده است.

در این شبیه‌سازی از مجموعه‌ی داده‌های کیوی [۱] استفاده می‌شود. این مجموعه شامل ۷ نمونه‌ی کیوی است که از هر کدام طیف مادون قرمز شامل ۵۰ طول موج مختلف عبور داده شده و پس از خروج از آن اندازه‌گیری می‌شود. این اندازه‌های خروجی به‌عنوان مقادیر متغیرهای توضیحی در نظر گرفته می‌شوند. متغیر پاسخ نیز میزان اسید موجود در کیوی می‌باشد که با استفاده از تجزیه‌ی شمیایی کیوی اندازه‌گیری خواهد شد.

۱-۴- نحوه‌ی شبیه‌سازی

متغیرهای توضیحی دارای توزیع نرمال چندمتغیره به‌صورت زیر هستند:

$$\mathbf{X} = (X_1, X_2, \ldots, X_p)' \sim \mathrm{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Gamma})$$

متغیر Y نیز با معادله‌ی رگرسیونی زیر ارائه خواهد شد:

$$Y = \alpha_0 + \boldsymbol{\alpha}'\mathbf{x} + \varepsilon \qquad (۸)$$

که در آن $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ مقدار نمونه‌ای $\mathbf{X}$ است. α_0 و $\boldsymbol{\alpha}'$ ثابت‌های نامعلوم هستند و $\varepsilon \sim \mathrm{N}(0, \sigma^2)$.

برای این‌که اثر افزایش تعداد متغیرهای توضیحی بر نحوه‌ی عملکرد روش‌های رگرسیونی بررسی شود از مدل‌هایی با ۶، ۲۰ و ۵۰ متغیر توضیحی استفاده خواهد شد. به این ترتیب که ابتدا به‌طور تصادفی ۲۰ متغیر توضیحی از بین ۵۰ متغیر انتخاب می‌شود و سپس از بین این ۲۰ متغیر توضیحی تعداد ۶ متغیر به‌طور تصادفی انتخاب می‌گردد.

۲-۴- محاسبه‌ی مقادیر پارامتر

برای محاسبه‌ی مقادیر پارامتر α_0 و $\boldsymbol{\alpha}'$ در مدل (۸) از مجموعه‌ی داده‌های کیوی استفاده می‌شود. ابتدا میانگین و ماتریس کوواریانس ۵۰ متغیر توضیحی محاسبه شده و به‌عنوان مقادیر $\boldsymbol{\mu}$ و $\boldsymbol{\Gamma}$ برای مدلی شامل ۵۰ متغیر توضیحی در نظر گرفته می‌شود. برای هر مدل با تعداد متغیر توضیحی کم‌تر یک زیرمجموعه‌ی تصادفی شامل ۶ و ۲۰ متغیر توضیحی از ۵۰ طول موج استخراج می‌شود. میزان اسید موجود در کیوی با استفاده از تجزیه‌ی شیمیایی اندازه‌گیری می‌شود و به‌عنوان مقادیر متغیر پاسخ هر بار روی یک مجموعه از این متغیرهای توضیحی رگرسیون می‌شود و ضرایب رگرسیونی براورد شده به‌عنوان مقادیر α_0 و $\boldsymbol{\alpha}'$ در مدل (۸) در نظر گرفته می‌شوند.

برای مجموعه‌ی داده‌های کیوی واریانس خطا، σ^2، حدود ۰/۱۵ به دست آمده است. اما چنین تصور می‌شود که PLS نسبت به روش‌های دیگر بیش‌تر به این پارامتر حساسیت دارد. بنا بر این مقادیر ۱، ۰/۵، ۰/۳، ۰/۱۵ و ۰/۰۵، σ^2 = نیز در شبیه‌سازی مورد آزمایش قرار می‌گیرند.

۳-۴- روش‌های رگرسیونی

در این شبیه‌سازی عملکرد چهار روش رگرسیونی از نظر توانایی پیش‌بینی مورد ارزیابی قرار می‌گیرد.

روش اول رگرسیون PLS است که در بخش دوم شرح داده شد. همچنین برای انتخاب تعداد مؤلفه‌ها در PLS از روش اعتبارسنجی متقابل استفاده می‌شود. برای این منظور ابتدا داده‌ها به پنج گروه تقسیم شده و هر بار یک گروه حذف می‌شود و از داده‌های چهار گروه دیگر برای ساختن مؤلفه‌ها و تعیین یک معادله‌ی پیش‌بینی برای متغیر پاسخ Y استفاده می‌شود. تابع زیان مورد استفاده در روش اعتبارسنجی متقابل به‌صورت زیر تعریف می‌شود:

$$\mathrm{RMSECV} = \sqrt{\sum_{i=1}^{n} (\hat{y}_{CV,i} - y_i)^2 / n}$$

که در آن $\hat{y}_{CV,i}$ پیش‌بینی i امین مشاهده است هنگامی که در مدل حضور ندارد. این تابع زیان جذر میانگین توان‌های دوم خطای پیش‌بینی اعتبارسنجی متقابل (Root Mean Squares Error of Cross Validation) نامیده می‌شود.

دومین روش، رگرسیون کم‌ترین توان‌های دوم معمولی (Ordinary Least Squares) است که به اختصار با OLS نمایش داده می‌شود. سومین روش، انتخاب متغیر پیش‌رو (Forward Variable Selection) است و به اختصار با FVS نمایش داده می‌شود. در این روش از آماره‌ی F جزئی برای ورود متغیر به مدل استفاده می‌گردد. اگر مقدار این آماره برای یک متغیر دلخواه از صدک ۹۵ام توزیع F یعنی ۴/۳۵ بیش‌تر شود آن متغیر وارد مدل خواهد شد.

آخرین روش مورد استفاده در این شبیه‌سازی رگرسیون مؤلفه‌های اصلی (Principal Component Regression) است که به اختصار با PCR نمایش داده می‌شود و مانند PLS براوردگرهایی اریب از ضرایب رگرسیونی تولید می‌کند. در این روش نیز مانند PLS از ترکیبات خطی متغیرهای توضیحی برای کاهش اثر مسئله‌ی هم‌خطی استفاده می‌گردد. این ترکیبات خطی که مؤلفه‌های اصلی نامیده می‌شوند از روش تحلیل مؤلفه‌های اصلی و با استفاده از بردارهای ویژه ماتریس کوواریانس متغیرهای توضیحی $\mathbf{X}$ به دست می‌آیند و به‌عنوان متغیرهای توضیحی جدید در رگرسیون مورد استفاده قرار

می‌گیرند. در این‌جا نیز برای انتخاب تعداد مؤلفه‌هایی که باید در مدل بمانند از روش FVS استفاده می‌گردد. به این ترتیب که اگر مقدار آماره‌ی F جزئی برای یک مؤلفه از صدک ۹۵ام توزیع F یعنی ۴/۳۵ بزرگ‌تر باشد آن مؤلفه وارد مدل خواهد شد.

۴-۴- تابع زیان و روش شبیه‌سازی

فرض کنید یک معادله‌ی پیش‌بینی که با استفاده از روش‌های ذکر شده در بخش قبل بنا شده به‌صورت $\hat{y} = \hat{\alpha}_{\circ} + \hat{\boldsymbol{\alpha}}'\mathbf{x}$ باشد. آن‌گاه با توجه به مدل (۸) داریم

$$y - \hat{y} = (\alpha_{\circ} - \hat{\alpha}\circ) + (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})'\mathbf{x} + \varepsilon$$

که در آن $\hat{\alpha}_{\circ}$ و $\hat{\boldsymbol{\alpha}}$ معلوم و توزیع $\mathbf{X}$ها و ε نیز معلوم است. متغیر تصادفی $y - \hat{y}$ را می‌توان به‌عنوان خطای حاصل از پیش‌بینی y توسط مشاهده‌ی آینده $\hat{y}$ در نظر گرفت. اگر از میانگین توان‌های دوم خطای پیش‌بینی y توسط $\hat{y}$ سهم خطای تصادفی $\sigma^{۲}$ کسر شود آن‌چه باقی می‌ماند زیانی است که ناشی از عدم دقت در براورد ضرایب رگرسیونی می‌باشد. این زیان به‌عنوان تابع زیان برای این شبیه‌سازی در نظر گرفته می‌شود و به‌صورت زیر تعریف می‌شود:

$$Loss = \{(\alpha_{\circ} - \hat{\alpha}_{\circ}) + (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})'\boldsymbol{\mu}\}^{۲} + (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})'\boldsymbol{\Gamma}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) \qquad (۹)$$

در این شبیه‌سازی اندازه‌ی مدل یعنی تعداد متغیرهای توضیحی برابر ۵۰، ۲۰، ۶ = p و واریانس خطای تصادفی نیز برابر ۱، ۰/۵، ۰/۳، ۰/۱۵، ۰/۰۵ = $\sigma^{۲}$ در نظر گرفته می‌شود و با استفاده از مقادیر پارامتر متناظر با اندازه‌ی مدل یک مجموعه نمونه از (p+۲۰) داده شبیه‌سازی می‌شود که هر داده شامل مقادیر X و Yها است.

از این نمونه و با استفاده از روش‌های رگرسیونی ذکر شده معادلات پیش‌بینی براورد شده و دقت این معادلات با استفاده از تابع زیان (۹) اندازه‌گیری می‌شود. این عمل ۵۰۰ بار برای هر اندازه مدل (p) و واریانس خطا ($\sigma^{۲}$) تکرار می‌شود و متوسط زیان برای هر روش رگرسیون محاسبه می‌گردد.

۵-۴- نتایج شبیه‌سازی

نتایج این شبیه‌سازی با توجه به اندازه‌های مختلف مدل و سطوح مختلف واریانس خطای تصادفی برای چهار روش رگرسیونی در جدول (۲) نمایش داده شده است. مقادیر داخل جدول مربوط به متوسط زیان حاصل از این روش‌های رگرسیونی در ۵۰۰ تکرار آزمایش می‌باشد. مقادیر داخل پرانتز نیز مربوط به متوسط تعداد مؤلفه‌ها یا متغیرهایی است که در این ۵۰۰ تکرار مورد استفاده قرار گرفته‌اند.

همان‌طور که مشاهده می‌شود زمانی که شش متغیر توضیحی در مدل حضور دارند روش FVS در سطوح مختلف واریانس خطا (غیر از سطح اول) دارای متوسط زیان بیش‌تری نسبت به روش‌های دیگر است. روش OLS نیز بعد از FVS دارای بیش‌ترین مقدار زیان است و این بدان معناست که در بین روش‌های مورد استفاده در این شبیه‌سازی معادلات ساخته شده توسط این دو روش کم‌ترین توان پیش‌بینی را دارند. این اختلاف عملکرد با افزایش واریانس خطا فاحش‌تر می‌شود. روش‌های اریب PLS و PCR عملکرد تقریباً یکسانی دارند اما با افزایش واریانس خطا روش PLS دارای متوسط زیان کم‌تری نسبت به بقیه‌ی روش‌ها می‌شود البته این اختلاف چندان زیاد نیست.

زمانی که تعداد متغیرهای توضیحی به ۲۰ افزایش می‌یابد OLS بدترین عملکرد را نسبت به بقیه‌ی روش‌ها پیدا می‌کند. در این حالت روش‌های FVS و PCR عملکرد یکسانی دارند. این در حالی است که PLS در تمام سطوح واریانس معادلاتی با زیان پیش‌بینی کم‌تر از بقیه روش‌ها بنا می‌کند و این بهبود در پیش‌بینی با افزایش واریانس خطا بیش‌تر احساس می‌گردد.

اما زمانی که ۵۰ متغیر توضیحی در مدل حضور دارند عملکرد OLS خیلی بدتر از قبل می‌شود. روش PCR نیز بعد از OLS دارای بیش‌ترین زیان پیش‌بینی است. در این حالت PLS در تمام سطوح بهترین عملکرد را دارد و این حالت با افزایش واریانس خطا بیش‌تر به چشم می‌خورد. روش FVS نسبت به قبل بهتر شده و در تمام سطوح واریانس عملکردی نزدیک به PLS دارد.

با توجه به مطالب ذکر شده می‌توان گفت: توان پیش‌بینی تمام روش‌ها با افزایش واریانس خطا کاهش می‌یابد اما میزان کاهش در روش‌های رگرسیونی اریب به مراتب کم‌تر از روش کم‌ترین توان‌های دوم معمولی است. همچنین زمانی که بین متغیرهای توضیحی هم‌خطی شدیدی وجود دارد روش‌های PLS و PCR عملکرد متفاوتی دارند. در

جدول۲- متوسط زیان چهار روش رگرسیونی با اندازه‌های مختلف مدل و سطوح متفاوت واریانس خطای تصادفی

اندازه مدل	واریانس خطای تصادفی	روش‌های رگرسیونی			
		OLS	FVS	PCR	PLS
۶	۰/۰۵	۰/۰۱۸۸	۰/۰۲۱(۱/۰۲)	۰/۰۲۲(۱/۸۱)	۰/۰۱۸(۴/۶۰)
۶	۰/۱۵	۰/۰۵۷	۰/۰۸۳(۰/۴۱)	۰/۰۴۲(۰/۸۰)	۰/۰۴۸(۴/۰۵)
۶	۰/۳	۰/۱۱۸	۰/۱۵۵(۰/۲۲)	۰/۰۷۳(۰/۶۲)	۰/۰۶۴(۳/۴۰)
۶	۰/۵	۰/۱۹۱	۰/۲۱۷(۰/۱۷)	۰/۰۹۸(۰/۵۰)	۰/۰۹۹(۲/۱۲)
۶	۱	۰/۳۸۸	۰/۴۹۲(۰/۱۴)	۰/۱۶۶(۰/۴۷)	۰/۱۵۷(۲/۰۱)
۲۰	۰/۰۵	۰/۰۵۸	۰/۰۶۲(۱/۹۶)	۰/۰۶۶(۷/۶۲)	۰/۰۷۰(۶/۷۲)
۲۰	۰/۱۵	۰/۱۷۸	۰/۱۷۷(۰/۹۷)	۰/۱۳۵(۴/۳۷)	۰/۰۹۸(۴/۴۹)
۲۰	۰/۳	۰/۳۴۸	۰/۲۱۴(۰/۵۰)	۰/۲۱۲(۳/۵۱)	۰/۱۱۴(۳/۲۸)
۲۰	۰/۵	۰/۵۹۰	۰/۳۲۷(۰/۴۱)	۰/۳۱۸(۲/۹۵)	۰/۱۴۹(۲/۷۰)
۲۰	۱	۱/۱۶۱	۰/۴۹۱(۰/۲۴)	۰/۵۰۵(۲/۱۳)	۰/۲۳۱(۲/۰۵)
۵۰	۰/۰۵	۰/۱۴۱	۰/۱۰۸(۳/۸۹)	۰/۱۴۰(۲۲/۷۵)	۰/۰۸۲(۱۲/۸۲)
۵۰	۰/۱۵	۰/۴۳۳	۰/۱۴۰(۱/۹۷)	۰/۳۴۱(۱۶/۶۷)	۰/۱۳۵(۱۱/۶۶)
۵۰	۰/۳	۰/۸۲۴	۰/۲۲۹(۰/۹۲)	۰/۵۸۲(۱۴/۰۷)	۰/۱۹۴(۱۰/۶۴)
۵۰	۰/۵	۱/۴۳۷	۰/۳۱۰(۰/۶۴)	۰/۹۱۱(۱۲/۰۶)	۰/۲۶۲(۹/۱۱)
۵۰	۱	۲/۷۲۰	۰/۴۶۰(۰/۳۹)	۱/۶۰۸(۱۰/۲۳)	۰/۳۸۷(۸/۴۲)

این حالت PLS از بقیه‌ی روش‌ها بهتر عمل می‌کند و نتایج پیش‌بینی معتبرتری نسبت به دیگر روش‌ها ارائه می‌دهد. در حالت کلی، این شبیه‌سازی مزیت استفاده از روش‌های رگرسیون اریب را در وضعیت‌هایی که واریانس خطای تصادفی زیاد و هم‌خطی شدید است نشان می‌دهد.

همان‌طور که گفته شد این شبیه‌سازی بر اساس شبیه‌سازی انجام شده در مرجع [۴] صورت گرفته است. گارسویت [۴] با استفاده از ۱۹۵ نمونه‌ی علوفه و عبور۵۰ طول موج طیف مادون قرمز از آن‌ها عملکرد روش PLS را با روش‌های رگرسیونی ذکر شده در بخش (۳-۴) مورد مقایسه قرار داد. او این مقایسه را در پنج سطح مختلف واریانس خطا با مقادیر ۱۰، ۷،۳، ۱ = σ^2 با استفاده از مدل‌هایی شامل ۵۰، ۲۰، ۸ = p

متغیر توضیحی انجام داد و با استفاده از شبیه‌سازی تعداد ($p+۴۰$) داده در ۵۰۰ تکرار آزمایش برای هر اندازه‌ی مدل و واریانس خطا، متوسط زیان حاصل از چهار روش رگرسیونی را مورد بررسی قرار داد. او با استفاده از شبیه‌سازی داده‌های علوفه نشان داد زمانی که تعداد متغیرهای توضیحی خیلی زیاد و واریانس خطای تصادفی نیز بزرگ است روش PLS عملکرد خوبی نسبت به روش‌های دیگر دارد و در این حالت معادلات پیش‌بینی بهتری نسبت به دیگر روش‌های رگرسیونی بنا می‌کند. بنا بر این می‌توان نتیجه گرفت که نتایج به دست آمده از شبیه‌سازی داده‌های علوفه و داده‌های کیوی تقریباً یکسان هستند.

مرجع‌ها

[۱] مقیمی، علی (۱۳۸۷). تعیین کیفیت درونی میوه‌ی کیوی به‌صورت غیر مخرب با استفاده از طیف‌سنجی مادون قرمز. پایان‌نامه‌ی کارشناسی ارشد، دانشگاه فردوسی مشهد.

[2] Dunn, W.J. and Ruhe, A. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *Siam J. Sci. Stat. Comput*. **5**, 735–743.

[3] Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

[4] Garthwaite, P.H. (1994). An interpretation of partial least squares. *J. Ame. Statist. Assoc*. **89**, 122–127.

[5] Garthwaite, P.H., Jolliffe, I.T. and Jones, B. (1995). *Statistical Inference*. Prentice–Hall.

[6] Helland, I.S. (1988). On the structure of partial least squares. *Communs Statist. Simuln*, **17**, 581–607.

[7] Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). *Introduction to Linear Regression Analysis*. Wiley, New York

[8] Naes, T. and Martens, H. (1985). Comparison of prediction methods for multicollinear data. *Commun. Statist. Simula. Computa*. **14**, 545–576.

[9] Wold, H. (1982). *Soft Modeling: The Basic Design and Some Extensions*. In Wold, H. and K.G. Joreskog, editors: *Systems Under Indirect Observations: Causality, Structure, Prediction*. Elsevier, Amsterdam.

امیر شاهینی
کارشناس ارشد آمار
نشانی: استان خراسان‌رضوی، مشهد، دانشگاه فردوسی مشهد، دانشکده‌ی علوم ریاضی، صندوق پستی ۱۱۵۹-۹۱۷۷۵
رایانشانی: ams_stat@yahoo.com

مجید سرمد
دکتری آمار
استان خراسان‌رضوی، مشهد، دانشگاه فردوسی مشهد، دانشکده‌ی علوم ریاضی، صندوق پستی ۱۱۵۹-۹۱۷۷۵
رایانشانی: sarmad@um.ac.ir