Revista de Matemática: Teoría y Aplicaciones
ISSN: 1409-2433
mta.cimpa@ucr.ac.cr
Universidad de Costa Rica
Costa Rica

Vega–Vilca, José Carlos; Guzmán, Josué
Regresión PLS y PCA como solución al problema de multicolinealidad en regresión múltiple
Revista de Matemática: Teoría y Aplicaciones, vol. 18, núm. 1, 2011, pp. 9-20
Universidad de Costa Rica
San José, Costa Rica

Revista de Matemática: Teoría y Aplicaciones 2011 **18**(1) : 9–20
cimpa – ucr issn: 1409-2433

REGRESIÓN PLS Y PCA COMO SOLUCIÓN AL PROBLEMA DE MULTICOLINEALIDAD EN REGRESIÓN MÚLTIPLE

José Carlos Vega–Vilca* Josué Guzmán†

Received: 18 Feb 2010; Revised: 5 Aug 2010; Accepted: 14 Aug 2010

Resumen

Se presentan y comparan las técnicas de regresión por componentes principales y la regresión por componentes desde mínimos cuadrados parciales, como solución al problema de multicolinealidad en regresión múltiple. Se ilustran las metodologías con ejemplos de aplicación en la que se observa la superioridad de la técnica por mínimos cuadrados parciales.

Palabras clave: análisis de componentes principales, mínimos cuadrados parciales, reducción de la dimensionalidad.

Abstract

We present and compare principal components regression and partial least squares regression, and their solution to the problem of multicollinearity. We illustrate the use of both techniques, and demonstrate the superiority of partial least squares.

Keywords: principal components analysis, partial least squares, dimensionality reduction.

Mathematics Subject Classification: 62H25, 62J02.

*Instituto de Estadística, Universidad de Puerto Rico – Recinto de Río Piedras, Puerto Rico. E-Mail: josecvega07@gmail.com

†Programa Doctoral de Administración de Empresas, Universidad del Turabo, Gurabo, Puerto Rico. E-Mail: jguzmanphd@gmail.com

1 Introducción

En la construcción de un modelo de regresión lineal múltiple se pueden presentar dos problemas: multicolinealidad y alta dimensionalidad de sus variables predictoras. En este trabajo se revisan dos metodologías relativamente similares y usadas en la solución de estos problemas: Regresión por Mínimos Cuadrados Parciales (PLS, por sus siglas en inglés) y Regresión por Análisis de Componentes Principales (PCA, por sus siglas en inglés). Ambos métodos transforman las variables predictoras en componentes ortogonales, los cuales representan la solución al problema de multicolinealidad y permiten hacer una reducción de la dimensionalidad del espacio de variables predictoras.

Frank y Friedman (1993) [1] afirman que la regresión PLS es fuertemente promocionada y usada por especialistas en Quimiometría, pero desconocido por estadísticos; mientras que regresión PCA es bastante conocido pero muy pocas veces recomendado por estadísticos.

El objetivo del presente trabajo es difundir la teoría y aplicación de la regresión PLS, muy usada en un área de la química llamada Quimiometría, para que pueda ser aplicada en toda disciplina que trabaja con datos caracterizados por muchas variables medidas sobre cada uno de pocos sujetos.

2 Multicolinealidad

La multicolinealidad describe la dependencia lineal entre las variables predictoras; es un problema que hace difícil cuantificar con precisión el efecto que cada variable predictora ejerce sobre la variable dependiente y puede ser determinada mediante el cálculo del Factor de Inflación de Varianza (VIF, por sus siglas en inglés) y por el número condición (η).

El VIF es un indicador de multicolinealidad específica de cada variable predictora del modelo. $VIF_j = \frac{1}{1-R_j^2}$ para $j = 1, 2, \cdots, p$; donde R_j^2 es el coeficiente de determinación de la regresión lineal de X_j respecto a las demás variables predictoras. Como regla general, si $VIF_j \geq 10$, entonces existe multicolinealidad.

El número condición es un indicador de multicolinealidad global de las predictoras del modelo. $\eta = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$; donde λ_{max} y λ_{min} son los autovalores máximo y mínimo de la matriz de correlaciones entre predictoras $\mathbf{R}$. También como regla general, si $\eta \geq 30$, entonces existe multicolinealidad.

3 Regresión por Análisis de Componentes Principales (Regresión PCA)

La regresión por componentes principales es un método introducido por Massy (1965) [6] que aplica mínimos cuadrados sobre un conjunto de variables latentes llamadas componentes principales, obtenidas a partir de la matriz de correlación. Sea $\mathbf{X}$ de orden $n \times p$, la matriz de predictoras que al ser estandarizada por columnas origina la matriz $\mathfrak{X}$ de orden $n \times p$. La matriz de correlaciones entre predictoras está dada por $\mathbf{R} = (n-1)^{-1}\mathfrak{X}'\mathfrak{X}$, de orden $p \times p$. Usando descomposición espectral de una matriz simétrica se tiene que:

$$\mathbf{R} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}' \tag{1}$$

donde $\mathbf{\Gamma} = (\gamma_1, \cdots, \gamma_p)$ es una matriz ortogonal de orden $p \times p$, cada γ_i es llamado autovector y tiene norma 1. La matriz $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_p)$ es diagonal de orden $p \times p$; los λ_i son llamados autovalores y $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Por ortogonalidad de la matriz $\mathbf{\Gamma}$, la expresión (1) puede ser escrita como:

$$\mathbf{\Gamma}'\mathbf{R}\mathbf{\Gamma} = \mathbf{\Lambda} \tag{2}$$

se puede verificar la siguiente equivalencia para $i, j = 1, \cdots, p$

$$\gamma_i'\mathbf{R}\gamma_j = \begin{cases} \lambda_i & \text{si } i = j \\ 0 & \text{si } i \neq j. \end{cases} \tag{3}$$

La matriz de componentes principales $\mathbf{C}$ de orden $n \times p$, es obtenida transformando la matriz $\mathfrak{X}$, de la siguiente manera:

$$\begin{aligned} \mathbf{C} &= \mathfrak{X}\mathbf{\Gamma} \\ &= \mathfrak{X}(\gamma_1, \cdots, \gamma_p) \\ &= (\mathfrak{X}\gamma_1, \cdots, \mathfrak{X}\gamma_p) \end{aligned} \tag{4}$$

Cada $\mathfrak{X}\gamma_i$, para $i = 1, \cdots, p$ es llamada componente principal. De (3) se concluye que las componentes principales son ortogonales entre sí.

3.1 Fundamento del Análisis de Componentes Principales

La idea es maximizar la varianza del componente principal $\mathfrak{X}\gamma$ sujeto a que el autovector γ, satisfaga la restricción de ortogonalidad: $\gamma'\gamma = 1$

$$\begin{aligned} var(\mathfrak{X}\gamma) &= \gamma' var(\mathfrak{X})\gamma \\ &= \gamma'[(n-1)^{-1}\mathfrak{X}'\mathfrak{X}]\gamma \\ &= \gamma'\mathbf{R}\gamma. \end{aligned} \tag{5}$$

La función lagrangiana ϕ es usada para maximizar la varianza del componente principal, sujeto a la restricción de ortogonalidad del vector γ

$$\phi = \gamma' \mathbf{R} \gamma - \lambda(\gamma'\gamma - 1). \tag{6}$$

La maximización de ϕ determina al vector γ que maximiza la varianza del componente principal dada en la expresión (5). Derivando (6) con respecto a γ e igualando a cero, se tiene que:

$$\frac{\partial \phi}{\partial \gamma} = 2\mathbf{R}\gamma - 2\lambda\gamma = 0$$

$$\Rightarrow \ \mathbf{R}\gamma = \lambda\gamma. \tag{7}$$

El cumplimiento de la igualdad de la expresión (7) implica que λ y γ son el autovalor y el autovector, respectivamente, de la matriz de correlaciones $\mathbf{R}$. La relación entre λ y γ es unívoca, es decir a cada autovalor le corresponde un autovector; esto es demostrado en Mardia et al. (1997) [5].

4 Regresión por Mínimos Cuadrados Parciales (Regresión PLS)

La regresión PLS, fue introducida por H. Wold (1975) [9], para ser aplicada en ciencias económicas y sociales. Sin embargo, gracias a las contribuciones de su hijo Svante Wold, ha ganado popularidad en Quimiometría, en donde se analizan datos que se caracterizan por muchas variables predictoras, con problemas de multicolinealidad, y pocas unidades experimentales (observaciones o casos) en estudio.

La idea motivadora de PLS fue heurística, por ello algunas de sus propiedades son todavía desconocidas a pesar de los progresos alcanzados por Helland (1988) [3], Hoskulsson (1988) [4], Stone y Brooks (1990) [7] y otros. La metodología PLS generaliza y combina características del Análisis de Componentes Principales y Análisis de Regresión Múltiple. La demanda por esta metodología y la evidencia de que trabaja bien, van en aumento y así, la metodología PLS está siendo aplicada en muchas ramas de la ciencia.

En general, la regresión PLS consta de dos pasos fundamentales. Primero, transforma a la matriz de predictoras $\mathbf{X}$ de orden $n \times p$, con ayuda del vector de respuestas $\mathbf{Y}$ de orden $n \times 1$, en una matriz de componentes o variables latentes no correlacionados, $\mathbf{T}=(\mathbf{T}_1, \cdots, \mathbf{T}_p)$ de orden $n \times p$, llamados componentes PLS; esto contrasta con el análisis de componentes principales en el cual los componentes son obtenidos usando sólo la matriz

de predictoras $\mathbf{X}$. Segundo, calcula el modelo de regresión estimado usando el vector de respuestas original y como predictoras, los componentes PLS.

La reducción de la dimensionalidad puede ser aplicada directamente sobre los componentes ya que estos son ortogonales. El número de componentes necesarios para el análisis de regresión debe ser mucho menor que el número de predictoras.

4.1 Algoritmo de la Regresión PLS

El siguiente algoritmo es adaptado de Garthwaite (1994) [2] y Trygg (2001) [8]. La entrada de datos corresponde a la matriz de predictoras $\mathbf{X}(n \times p)$ y el vector respuesta $\mathbf{Y}(n \times 1)$, los cuales han sido estandarizadas por columnas.

Algoritmo PLS
1. Entrada : $\mathbf{X}(0)$, $\mathbf{Y}(0)$
2. Para $h = 1$ hasta p
3. $\mathbf{w} = cov(\mathbf{Y}, \mathbf{X})$: normalizar $\mathbf{w}$
4. $\mathbf{T}_h = \mathbf{X}\mathbf{w}$
5. $v = (\mathbf{T}_h'\mathbf{Y})/(\mathbf{T}_h'\mathbf{T}_h)$
6. $\mathbf{b} = (\mathbf{T}_h'\mathbf{X})/(\mathbf{T}_h'\mathbf{T}_h)$
7. $\mathbf{X}(h) = \mathbf{X}(h-1) - \mathbf{T}_h\mathbf{b}$
8. $\mathbf{Y}(h) = \mathbf{Y}(h-1) - \mathbf{T}_h v$
9. Próximo h

4.1.1 Descripción del algoritmo PLS

La matriz de datos puede ser escrita como $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_p)$, donde $\mathbf{X}_1, \cdots, \mathbf{X}_p$ son las columnas de la matriz $\mathbf{X}$. A continuación se describen los principales pasos de este algoritmo.

Paso 1 *Son los datos iniciales, estandarizados por columnas.*

Paso 2 *Se da inicio al cálculo del primer componente PLS*

Paso 3 *Se calcula el vector* $\mathbf{w} = (w_1, \cdots, w_p)'$*; cada elemento* w_i *es la covarianza de la variable respuesta con cada predictora. Finalmente* $\mathbf{w}$ *es normalizado a la unidad.*

Paso 4 *Se calcula el componente PLS,*

$$\mathbf{T}_h = \mathbf{X}\mathbf{w} = (\mathbf{X}_1, \cdots, \mathbf{X}_p) \cdot (w_1, \cdots, w_p)'$$

Paso 5 *Se calcula el coeficiente de regresión simple de la variable respuesta sobre el vector componente PLS, calculado en el paso anterior.*

Paso 6 *Se calcula el vector* $\mathbf{b} = (b_1, \cdots, b_p)$ *; cada elemento de este vector es el coeficiente de regresión simple de* $\mathbf{X}_i$ *sobre* $\mathbf{T}_h$

Paso 7 *Se actualiza la matriz de predictoras.*

Paso 8 *Se actualiza el vector de respuestas.*

Paso 9 *Se calcula el próximo componente PLS, a partir del Paso 3.*

4.2 Fundamento del PLS

La idea es maximizar el cuadrado de la covarianza entre el componente $\mathbf{T}_h = \mathbf{X}\mathbf{w}$, y la variable respuesta $\mathbf{Y}$, sujeto a $\mathbf{w}'\mathbf{w} = 1$. El componente $\mathbf{T}_h$ está definido como una combinación lineal de las predictoras, tal que $\mathbf{w} \neq \mathbf{0}$. Sea la matriz $\mathbf{A}$ de orden $p \times 1$, el vector de covarianzas de $\mathbf{X}$ e $\mathbf{Y}$. El análisis de regresión establece la dependencia de $\mathbf{Y}$ sobre las predictoras $\mathbf{X}$, por lo que $\mathbf{A} \neq \mathbf{0}$

$$\begin{aligned}[cov(\mathbf{X}\mathbf{w}, \mathbf{Y})]^2 &= [\mathbf{w}' cov(\mathbf{X}, \mathbf{Y})]^2 \\ &= [\mathbf{w}'\mathbf{A}]^2 \\ &= \mathbf{w}'\mathbf{A}\mathbf{A}'\mathbf{w}. \end{aligned} \tag{8}$$

La función lagrangiana ϕ usada para maximizar el cuadrado de la covarianza entre el componente $\mathbf{T}_h = \mathbf{X}\mathbf{w}$ y la variable respuesta $\mathbf{Y}$, sujeta a la restricción de ortogonalidad del vector $\mathbf{w}$, es:

$$\phi = \mathbf{w}'\mathbf{A}\mathbf{A}'\mathbf{w} - \lambda(\mathbf{w}'\mathbf{w} - 1).$$

Derivando ϕ respecto de $\mathbf{w}$, e igualando a cero, se tiene

$$\frac{\partial \phi}{\partial \mathbf{w}} = 2\mathbf{A}\mathbf{A}'\mathbf{w} - 2\lambda\mathbf{w} = \mathbf{0}$$

$$\Rightarrow \quad \mathbf{A}\mathbf{A}'\mathbf{w} = \lambda\mathbf{w}. \tag{9}$$

La igualdad de la expresión (9) implica que λ y $\mathbf{w}$ son el autovalor y el autovector de la matriz $\mathbf{A}\mathbf{A}'$, respectivamente.

Al multiplicar la expresión (9) por $\mathbf{w}'$, desde la izquierda

$$\mathbf{w}'\mathbf{A}\mathbf{A}'\mathbf{w} = \lambda. \tag{10}$$

Al multiplicar la expresión (9) por $\mathbf{A}'$, desde la izquierda

$$\mathbf{A}'\mathbf{A}\mathbf{A}'\mathbf{w} = \lambda\mathbf{A}'\mathbf{w}$$
$$(\mathbf{A}'\mathbf{A} - \lambda)\mathbf{A}'\mathbf{w} = 0$$

$$\mathbf{A}'\mathbf{A} - \lambda = 0 \quad \text{ó} \quad \mathbf{A}'\mathbf{w} = 0. \tag{11}$$

De la expresión (11) $\mathbf{A}'\mathbf{w}$ no puede ser cero, ya que se está buscando maximizarla, entonces $\mathbf{A}'\mathbf{A} - \lambda = 0$, de donde se obtiene la siguiente expresión

$$\lambda = \mathbf{A}'\mathbf{A} = \|\mathbf{A}\|^2. \tag{12}$$

De la expresión anterior $\lambda^2 = (\mathbf{A}'\mathbf{A})(\mathbf{A}'\mathbf{A}) = \lambda\|\mathbf{A}\|^2$, entonces:

$$\mathbf{A}'\mathbf{A}\mathbf{A}'\mathbf{A} = \lambda\|\mathbf{A}\|^2$$

$$\frac{\mathbf{A}'}{\|\mathbf{A}\|}\mathbf{A}\mathbf{A}'\frac{\mathbf{A}}{\|\mathbf{A}\|} = \lambda. \tag{13}$$

De (10) y (13), se puede reconocer que el vector $\mathbf{w}$ que maximiza al cuadrado de la covarianza del componente PLS y el vector de respuestas, es el vector de covarianzas normalizado

$$\mathbf{w} = \frac{\mathbf{A}}{\|\mathbf{A}\|} = \frac{\mathbf{X}'\mathbf{Y}}{\|\mathbf{X}'\mathbf{Y}\|}. \tag{14}$$

4.3 Propiedades observadas en Regresión PLS

Sea $\mathbb{U}$ un vector columna de unos, de dimensión n. Sean $\mathbf{X}(0)$ y $\mathbf{Y}(0)$ la matriz de predictoras y el vector de respuestas respectivamente, de datos iniciales estandarizados por columnas, por lo tanto se cumple: $\mathbf{X}'(0)\mathbb{U} = \mathbf{0}_{p\times 1}$ y $\mathbf{Y}'(0)\mathbb{U} = 0$. Además se cumplen las siguientes propiedades, las cuales pueden ser fácilmente probadas:

Propiedad 1 *El h-ésimo componente* $\mathbf{T}_h$*, siempre está centrado, es decir la suma de sus elementos es cero.* $\mathbf{T}_h'\mathbb{U} = 0$.

Propiedad 2 *La matriz de predictoras siempre está centrada en cualquier iteración, es decir la suma de cada una de sus columnas es cero.* $\mathbf{X}'(h)\mathbb{U} = \mathbf{0}_{p\times 1}$.

Propiedad 3 *El vector de respuestas siempre está centrado, es decir la suma de sus elementos es cero.* $\mathbf{Y}'(h)\mathbb{U} = 0$.

Propiedad 4 *En la h-ésima iteración, el componente* $\mathbf{T}_h$ *es ortogonal con cada una de las columnas de la matriz de predictoras.* $\mathbf{T}_h'\mathbf{X}(h) = \mathbf{0}_{1\times p}$.

Propiedad 5 *En la h-ésima iteración, se cumple que el componente* $\mathbf{T}_h$ *es ortogonal con el vector de respuestas.* $\mathbf{T}_h'\mathbf{Y}(h) = 0$.

Propiedad 6 *Cada par de componentes es ortogonal. Sean dos componentes* $\mathbf{T}_k$ *y* $\mathbf{T}_\ell$*, se cumple:* $\mathbf{T}_k'\mathbf{T}_\ell = 0$.

4.4 Matriz de transformación a componentes PLS

En análisis de componentes principales, la matriz que transforma las variables predictoras en componentes principales, es la matriz ortogonal $\mathbf{\Gamma}$, dada en la expresión (4). En análisis PLS, la matriz que transforma las variables predictoras en componentes PLS, es la matriz $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_p)$ de orden $p \times p$, la cual es hallada iterativamente de la siguiente manera:

$$\begin{aligned} \mathbf{z}_1 &= \mathbf{w}(1) \\ \mathbf{z}_h &= \left[\mathbf{I} - \sum_{j=1}^{h-1} \mathbf{z}_j \mathbf{b}(j)\right] \mathbf{w}(h) \quad ; \quad h > 1. \end{aligned} \tag{15}$$

Donde $\mathbf{I}$ es la matriz identidad, entonces la matriz de componentes PLS, se halla como en la siguiente expresión

$$\begin{aligned} \mathbf{T} &= \mathbf{X}(0)\mathbf{Z} \\ &= \mathbf{X}(0)(\mathbf{z}_1, \cdots, \mathbf{z}_p) \\ &= [\mathbf{X}(0)\mathbf{z}_1, \cdots, \mathbf{X}(0)\mathbf{z}_p]. \end{aligned} \tag{16}$$

Donde $\mathbf{T} = (\mathbf{T}_1, \cdots, \mathbf{T}_p)$, es la matriz de componentes PLS y $\mathbf{X}(0)$ es la matriz de predictoras de datos iniciales, estandarizada por columnas.

5 Selección del número de componentes

En análisis de regresión múltiple hay muchos criterios para seleccionar el mejor modelo de regresión. Se usó la suma de cuadrados de errores de predicción (PRESS, por sus siglas en inglés). Se estimaron los modelos de regresión de la variable dependiente y los *h-primeros* componentes principales y componentes PLS, se calculó el PRESS(h) de ambos modelos. El número óptimo de componentes fue determinado por la siguiente regla sugerida:

$$h* = min\{h > 1 : \text{PRESS}(h+1) - \text{PRESS}(h) > 0\}. \tag{17}$$

6 Aplicaciones

Para la aplicación se utilizó la base de datos $fat.R$, la cual está disponible en la librería UsingR, del programa R. Esta base de datos consta de 252 casos, 17 variables predictoras y una variable respuesta. La variable respuesta es Y: $body.fat$ y las 17 predictoras son: X_1: $body.fat.siri$, X_2: $density$, X_3: age, X_4: $weight$, X_5: $height$, X_6: BMI, X_7: $ffweight$, X_8: $neck$, X_9: $chest$, X_{10}: $abdomen$, X_{11}: hip, X_{12}: $thigh$, X_{13}: $knee$, X_{14}: $ankle$, X_{15}: $bicep$, X_{16}: $forearm$, X_{17}: $wrist$.

6.1 Aplicación 1

Se analiza la base de datos completa, el objetivo es detectar y eliminar los problemas de multicolinealidad; sobre todo, reducir la dimensionalidad. Se detectó problemas de multicolinealidad; los valores VIF, en 8 de las 17 variables predictoras, está en el rango de 11.3 a 97.6, lo cual está de acuerdo con el alto valor del número condición, $\eta = 43.317$.

La matriz de predictoras se transformó en componentes principales y componentes PLS, para eliminar la multicolinealidad. Se calculó el PRESS de la regresión PCA y PLS, respectivamente; estos resultados son mostrados en las Tablas 1 y 2. La regla dada por la expresión (17) sugiere que la regresión PCA sea con 6 componentes y la regresión PLS sea con 7 componentes.

Con fines de comparación, si ambos modelos de regresión PLS y PCA fuesen estimados con 6 componentes, los valores PRESS serían 88.31 y 266.54, respectivamente. Estos resultados confirman que el modelo PLS supera al modelo PCA.

8115.21(1)	1691.53(2)	1645.43(3)	556.04(4)	527.22(5)
266.54(6)	304.78(7)	313.61(8)	221.53(9)	223.56(10)
221.48(11)	219.31(12)	129.66(13)	195.46(14)	192.45(15)
78.51(16)	12.25(17)			

(.) número de componentes.

Tabla 1: PRESS con regresión PCA.

5135.62(1)	775.22(2)	162.43(3)	139.72(4)	103.97(5)	88.31(6)
61.58(7)	72.38(8)	35.63(9)	25.14(10)	18.26(11)	13.81(12)
13.48(13)	12.19(14)	12.68(15)	11.72(16)	12.25(17)	

(.) número de componentes.

Tabla 2: PRESS con regresión PLS.

6.2 Aplicación 2

Se presenta una muestra aleatoria de 15 casos extraídos al azar desde la base datos de la aplicación anterior, esto produce una matriz de variables predictoras de orden 15×17. Los casos son los siguientes: 226, 206, 117, 76, 41, 136, 121, 120, 130, 107, 214, 156, 69, 91, 73. En esta aplicación se ilustra el potencial de la regresión PLS, ya que el número de variables predictoras es mayor que el número de casos o de sujetos observados. No es posible la regresión por Mínimos Cuadrados Ordinarios, debido a las matrices singulares que se forman en los cálculos intermedios.

El cálculo del PRESS, se muestra en la Tabla 3. La regla de la expresión (17) sugieren el uso de 6 componentes PLS. El modelo de regresión estimado usando componentes PLS es presentado en la Tabla 4.

251.38(1)	83.29(2)	25.61(3)	7.79(4)	6.43(5)	**2.84(6)**
8.86(7)	22.41(8)	7.70(9)	1.94(10)	6.46(11)	29.55(12)
27.29(13)	27.29(14)	27.29(15)	27.29(16)	27.29(17)	

(.) número de componentes.

Tabla 3: PRESS con regresión PLS.

Componentes	Coef.	SE Coef.	t-cal	p-valor	VIF
Constante	18.533	0.0798	232.35	0.000	
pls1	1.661	0.0249	66.82	0.000	1.0
pls2	2.545	0.0622	40.93	0.000	1.0
pls3	1.290	0.0858	15.04	0.000	1.0
pls4	1.085	0.1069	10.15	0.000	1.0
pls5	0.378	0.1066	3.54	0.008	1.0
pls6	0.406	0.1628	2.49	0.037	1.0

PRESS = 2.8431

Tabla 4: Regresión estimada con componentes PLS.

6.2.1 Predicción

Con el fin de evaluar la predicción del modelo de regresión estimado, dado en la Tabla 4, se seleccionaron aleatoriamente 5 casos: 47, 118, 35, 92, 229, de la base de datos mencionada en la aplicación 1. Las predictoras de estos 5 casos fueron estandarizados usando la media y desviación estándar de las variables predictoras de los 15 casos en el estudio mencionado en aplicación 2 y posteriormente fueron transformados a componentes PLS,

	observaciones transformadas					
caso	pls1	pls2	pls3	pls4	pls5	pls6
47	-4.437	0.685	-0.773	-1.138	0.417	0.635
118	-0.093	-1.534	-0.083	0.656	-0.493	-0.155
35	6.889	-0.195	0.990	0.328	0.514	0.935
92	0.471	-0.192	-0.266	-0.033	0.435	0.833
229	-0.287	-0.595	-0.966	0.430	-0.337	0.039

Tabla 5: Observaciones transformadas a componentes PLS.

caso	Verdadero valor	Predicción	Intervalo de Predicción (95%)
47	11.2	11.092	10.206 – 11.978
118	14.1	14.830	14.034 – 15.627
35	31.1	31.684	30.745 – 32.624
92	18.1	18.949	18.139 – 19.758
229	15.0	15.653	14.876 – 16.430

Tabla 6: Predicciones.

multiplicando la matriz estandarizada de orden 5×17 por la matriz de transformación $\mathbf{Z}$, de orden 17×6, desde las expresiones (15) y (16). La Tabla 5, presenta los componentes PLS usados para obtener las respectivas predicciones desde el modelo de regresión en la Tabla 4. Las predicciones se presentan en la Tabla 6.

7 Conclusiones

- Los resultados del análisis de datos verifican la eficiencia de la regresión PLS sobre la regresión PCA, hecho que fue probado analíticamente por Frank y Friedman (1993) [1].

- Cuando se fija el mismo número de componentes para estimar los modelo de regresión PLS y PCA, se observa que el valor PRESS siempre es menor para el modelo PLS.

- Cuando el número de predictoras es mucho mayor que el número de observaciones o casos, el modelo de regresión PLS puede ser estimado eficientemente.

- La programación del algoritmo PLS, en lenguaje R, resultó muy eficiente para el análisis de datos que ilustran este trabajo.

Referencias

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

[2] Garthwaite, P.H. (1994) "An interpretation of partial least square regression", *Journal of the American Statistical Association* **89**(425): 122–127.

[3] Helland, I. (1988) "On the structure of partial least squares regression", *Communications in Statistics, Simulation and Computation*, **17**(2): 581–607.

[4] Hoskulsson, A. (1988) "PLS regression methods", *Chemometrics*, **2**: 211–228.

[5] Mardia, K.V.; Kent, J.T.; Bibby, J.M. (1997) *Multivariate Analysis*. Academic Press, London.

[6] Massy, W.F. (1965) "Principal Components Regression in Exploratory Statistical Research", *Journal of the American Statistical Association*, **60**: 234–246.

[7] Stone, M.; Brooks, R.J. (1990) "Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression", *Journal of the Royal Statistical Society* **52**: 237–269.

[8] Trygg, J. (2001) *Parsimonious Multivariate Models*. PhD Thesis, Umea University, Research Group for Chemometrics Department of Chemistry.

[9] Wold, H. (1975) "Soft modeling by latent variables; the nonlinear iterative partial least square approach", *Perspectives in Probability and Statistics*, Papers in Honour of M.S. Bartlett.