REVISTA DE LA FACULTAD DE CIENCIAS ECONÓMICAS

RESULTADOS ÓPTIMOS DEMANDAN HERRAMIENTAS DE MAYOR PRECISIÓN EL APORTE DE MÍNIMOS CUADRADOS PARCIALES (PLS)

ANA MARIA LEGATO
FACULTAD DE CIENCIAS ECONÓMICAS - UNICEN – CEA

ARGENTINA

ALDO HERNAN ALONSO
FACULTAD DE CIENCIAS ECONÓMICAS - UNLP – UNICEN

ARGENTINA

"El conocimiento científico es la antítesis del dogmatismo y nada puede contribuir más a impulsar el desarrollo. Para el dogma y los dogmáticos cualquier tiempo pasado fue mejor. Para la ciencia la razón de su existencia es la innovación y la necesidad de explorar."
Eduardo Punset

Recibido: 18/12/2012
Aceptado: 08/08/2013

RESUMEN

La regresión PLS (Partial Least Squares) es un método estadístico multivariante recientemente generalizado. Combina y generaliza conceptos de análisis de Componentes Principales y de análisis de Regresión Lineal Múltiple y resulta particularmente útil cuando se desea predecir un conjunto de variables dependientes (Y) desde un conjunto (relativamente grande y posiblemente correlacionadas) de variables predictoras (X). También resuelve con propiedad el problema de multicolinealidad, que generalmente se supera eliminando las variables que la causan o transformándolas, solución aplicable si la permanencia del set de variables X no es requerida, o sea cuando necesidades de explicación y predicción no inhiban tal procedimiento. Es apto asimismo cuando el problema requiere considerar relaciones múltiples y cruzadas, y que todas ellas se den simultáneamente o cuando existen variables que no se puedan medir directamente (no observables) no obstante ser necesarias para desarrollar la teoría.

El presente trabajo considera específicamente esta metodología, la describe e interpreta en su concepción y hace explícito su potencial aporte a través de su aplicación a dos casos simplificados[1] que permiten comparar los resultados con los obtenidos mediante el empleo de otra técnica.

Palabras clave: regresión PLS – análisis de Componentes Principales – regresión múltiple – multicolinealidad.

ABSTRACT

PLS regression (Partial Least Squares) is a multivariable statistical method that has been recently generalized. It combines and generalizes concepts of the analysis of Principal Components and the analysis of Multiple Linear Regression. It has proved to be particu-larly useful to predict a set of dependent variables (Y) from a quite big set of possibly correlated independent variables (X). It is also appropriate to solve the problem of mul-ticolinearity, a problem that is generally overcome by eliminating the variables that cause this phenomenon or even by transforming these variables, a solution that can be applied if the permanence of the X variables set is not required, that is to say, when the need of explanation and prediction does not inhibit such a procedure. It is also suitable when the problem requires to consider multiple and crossed relations, all of them occurring simul-taneously or, when there are variables that cannot be measured directly (not noticeable) even if they are necessary to develop the theory.

[1] No se consideran variables no observables, en la situación atendida en los ejemplos resueltos.

This paper considers especially this methodology; it describes it and interprets it from its very conception. Its potential input is explained by means of its application to two simplified cases that allow us to compare the results with others obtained by using another technique.

Key Words: PLS regression – analysis of Principal Component – multiple regression – multicolinearity.

1. INTRODUCCIÓN

Cuando la investigación demanda trabajar con numerosas variables y más aún si existe interacción entre ellas, resulta imprescindible recurrir a modelos de análisis multivariado. Por su intermedio se logra determinar con mayor precisión la resultante del comportamiento del conjunto de variables en juego a partir del aprovechamiento óptimo de la información contenida en los datos disponibles.

Si se pretende explicar una variable en función de otras, o sea cuando se busca determinar el comportamiento de una variable dependiente ("a explicar" o endógena) en función de más de una variable independiente (explicativas, exógenas o predictoras), uno de los métodos más difundidos es el Análisis de Regresión Lineal Múltiple (RLM), que recurre al método de Mínimos Cuadrados Ordinarios (MCO) o al de Mínimos Cuadrados Ponderados para la estimación. En todos los casos, las estimaciones deben resultar interpretables en el contexto real del problema, para lo cual, los coeficientes de regresión del modelo deben contener signos coincidentes con el de la correlación lineal entre la variable a explicar y las explicativas.

También, suele ocurrir que el número de variables explicativas sea muy numeroso, en cuyo caso el modelo pretendido puede ser de fatigosa obtención y de compleja interpretación, lo que hace procedente aplicar alguna técnica de reducción de dimensionalidad, como por ejemplo Análisis de Componentes Principales (ACP). Por su intermedio, se reemplazan algunas de las variables exógenas por su mejor combinación lineal, para entonces obtener una regresión sobre las nuevas variables generadas, lo que elimina eventuales problemas de multicolinealidad.

En algunos casos, puede suceder que en el conjunto de variables que intervienen en el problema, algunas deban reconocer el rol de variables explicativas en determinadas cuestiones mientras en otras deben ser las variables a explicar. En tales casos, la metodología adecuada puede ser Sistema de Ecuaciones Simultáneas, es decir, un modelo multiecuacional. Pero esa simultaneidad del comportamiento de las ecuaciones de un modelo añade nuevas dificultades, al menos en dos aspectos (Pulido & López, 1999): por un lado, la estimación de los parámetros del modelo puede que haga exigible procedimientos más complejos para poder recoger los efectos de esta simultaneidad y, por el otro, plantee la necesidad de la resolución conjunta de

un modelo simultáneo. Al respecto, la resolución matemática del problema no es compleja tratándose de un reducido número de ecuaciones lineales, pero se complica al crecer dicho número. Los múltiples procedimientos de estimación disponible, para este problema, pueden agruparse en dos grandes bloques (Pulido, 1987): *métodos de información completa y métodos de información limitada*, en lugar de su forma más inmediata o directa, que consiste en estimar cada una de las ecuaciones del modelo, como si fuera un modelo de ecuación única.

Otra situación emerge si, además de lo expuesto, se presentan relaciones múltiples y cruzadas, si existen variables que se reconocen necesarias en teoría pero que en la práctica no se pueden medir directamente (son "no observables" o latentes), o si la totalidad de las variables planificadas según se lo requiere en teoría, deban necesariamente permanecer en el modelo aunque no cumplan con algunas de las condiciones requeridas por la RLM o, simplemente cuando el número de variables supera el tamaño de la muestra. Con el objetivo de satisfacer las demandas planteadas en tales casos, han surgido en los últimos años nuevas técnicas estadísticas que, si bien se fundamentan en el análisis de regresión, recurren a otros algoritmos.

El presente trabajo considera, específicamente, el de Regresión por Mínimos Cuadrados Parciales – Partial Least Squares (PLS). Se lo describe e interpreta en su concepción y se hace explícito su potencial aporte a través de su aplicación a un caso simplificado que permite comparar resultados con los obtenidos con otra técnica.

El modelo de regresión PLS es usado en diversas disciplinas tales como química, economía, medicina, psicología y agronomía, entre otras ciencias que requieren modelos lineales con un gran número de predictores.

2. BREVE DESCRIPCIÓN CONCEPTUAL DE LOS MÉTODOS CONSIDERADOS

En la construcción de un modelo de regresión lineal múltiple basado en una matriz de datos **X**, de orden $n \times p$, (n es el número de observaciones y p las características observadas) se pueden presentar varios problemas, entre los cuales merece destacarse la multicolinealidad y alta dimensionalidad de sus variables predictoras. En esta presentación se revisan dos metodologías que siendo relativamente similares suelen ser usadas en la solución de dichos problemas: Regresión por Componentes Principales y Regresión por Mínimos Cuadrados Parciales, lo que se hará luego de recordar tanto el RLM y como el ACP.

Ambos métodos transforman las variables predictoras en variables artificiales, no observables originariamente, denominadas variables latentes o componentes principales, las cuales son ortogonales y permiten hacer una reducción de la dimensionalidad del espacio de variables predictoras[2]. Posteriormente, usando solamente las variables latentes se construye

[2] No es objeto del presente trabajo el Análisis de Correlación Canónica

el modelo de regresión estimado.

2.1 SOBRE REGRESIÓN LINEAL MÚLTIPLE

En su forma más simple, un modelo de regresión lineal basado en una matriz de datos **X**, de orden *n×p* y un vector **Y**, de orden *nx1*, expresa la relación (lineal) entre una variable independiente Y y un conjunto de predictoras X, de tal modo que:

$$Y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots + \beta_k x_{ki} + \mu_i$$

Donde β_o es el coeficiente de regresión que indica la ordenada al origen, en tanto que los β_i son los coeficientes de regresión (para las variables 1 a **p**) calculados en función de los datos estimados bajo los supuestos necesarios[3].

Entonces, se podría estimar (prever) el peso de una persona como función de su altura y del sexo, o las ventas de un producto en función del precio, calidad, promociones, y antigüedad de los vendedores, por ejemplo.

La eficacia del modelo obtenido se mide evaluando generalmente dos parámetros:

1. El ajustamiento, que mide la diferencia entre la respuesta proporcionada por el modelo y la respuesta experimental utilizada para crear el modelo, lo cual indica la medida en que el modelo encontrado se adapta a las observaciones reales.

2. La predicción, que indica la diferencia entre los puntos proporcionados por el modelo y los puntos experimentales obtenidos sucesivamente para la creación, representa la capacidad del modelo para predecir el fenómeno en cuestión.

No necesariamente un modelo que presenta un buen ajuste tiene una buena capacidad predictiva, por lo que se hace imprescindible validar cuidadosamente los modelos creados. Entre los problemas que se pueden presentar son comunes la multicolinealidad y alta dimensionalidad de sus variables predictoras y uno de los modos de solucionarlos es realizar la regresión sobre las Componentes Principales.

2.2 SOBRE REGRESIÓN POR COMPONENTES PRINCIPALES

La Regresión por Componentes Principales es un método que aplica mínimos cuadrados sobre un conjunto de variables artificiales llamadas precisamente Componentes Principales, obtenidas a partir –según las características de los datos– de la matriz de varianzas-covarianzas o de la matriz de correlación de las variables predictoras, una matriz designada **X**.

Esta metodología fue concebida por Pearson,(1901), quien publicó un trabajo sobre el

[3] A efectos de revisar los supuestos, se puede consultar, por ejemplo, Gujarati, (2004), *Elementos de Econometría*, pp. 54 - 63.

ajuste de un conjunto de puntos en un multiespacio a una línea o a un plano. Este enfoque fue retomado por Hotelling, (1933) quien fue el primero en formular el análisis por componentes principales tal como se ha difundido hasta el presente. Dicho enfoque se centra en el análisis de las componentes que sintetizan la mayor variabilidad del conjunto de puntos, lo que quizás explica el calificativo de "principal" (Plá, 1986). Por inspección de estos componentes que resumen la mayor proporción posible de la variabilidad total del conjunto de puntos, puede encontrarse un medio para clasificar o detectar las relaciones entre los puntos.

Desde la óptica del análisis de datos "a la francesa" (Langrand & Pinzón, 2009), se trataría de una técnica destinada a facilitar la lectura de la información contenida en grandes tablas de valores numéricos, sin plantear hipótesis de naturaleza estadística sobre los datos analizados, es decir, no se puede postular, sobre la base de conocimientos previos del universo en estudio, una estructura particular de las variables.

Desde sus orígenes, este análisis ha sido aplicado a situaciones muy variadas en psicología, medicina, geografía, agronomía y economía. Una de sus principales bondades reside en cuanto se desea conocer la relación entre los elementos de una población y se sospecha de que en dicha relación influye de manera desconocida un conjunto de variables o propiedades de los elementos.

2.2.1 GENERACIÓN DE LAS COMPONENTES PRINCIPALES

Las componentes principales tienen ciertas características "deseables", como:

• Las componentes principales *no están correlacionadas*, y si además, puede suponerse multinormalidad en los datos originales, son independientes.

• Cada Componente Principal sintetiza la *máxima variabilidad residual* contenida en los datos.

Al estudiar un conjunto de *n* individuos mediante *p* variables es posible encontrar nuevas variables denominadas Y_k, para k= 1,2,....., p que sean combinaciones lineales de las variables originales, x_j, especificadas en la matriz **X**, e imponer a este sistema ciertas condiciones que permitan satisfacer los objetivos del ACP.

Esto implica, encontrar (p x p) constantes tales que

$$Y_k = \sum_{j}^{p} wjk\, xj \text{ siendo } \quad k = 1,2,\ldots..,\ p$$

donde *wjk* es cada una de esas constantes.

En la expresión anterior puede observarse que debido a la sumatoria, en cada nueva variable Y_k intervienen todos los valores de las variables originales *xj*. El valor numérico de *wjk* indicará el grado de contribución que cada variable original aporta a la nueva variable definida

por la transformación lineal. Es posible que *wjk* tenga en algún caso particular el valor cero, lo cual indica que esta variable no influye en el valor de la nueva variable Y_k.

Ello significa que inicialmente se calculan las componentes principales de la matriz de predictores (no se consideran las endógenas) y se utilizan solamente las primeras componentes principales, o bien aquellas que contienen la máxima información. De este modo es posible reducir muchísimo el "rumor de fondo" (o sea todas aquellas oscilaciones instrumentales o típicas de casos reales, que ocasionan problemas de interpretación de los datos). Sobre las componentes principales (Y_k) así extraídas, se efectúa la regresión con el fin de obtener el modelo predictivo.

3. REGRESIÓN POR MÍNIMOS CUADRADOS PARCIALES (PARTIAL LEAST SQUARES, PLS)

El método fue introducido por Wold, H. (1975) para ser aplicado en ciencias económicas y sociales, para luego extenderse a otras disciplinas y, gracias a las contribuciones de su hijo Wold, S. (1983), también ha ganado popularidad en el área de la química conocida como *Chemometrics*, donde suelen darse casos en que se analizan datos caracterizados por muchas variables predictoras, con problemas de multicolinealidad y pocas unidades experimentales de estudio.

La idea motivadora de PLS fue heurística, por este motivo algunas de sus propiedades pueden ser aún inadvertidas a pesar de los progresos alcanzados por Helland, (1988) y Höskuldsson, (1988), entre otros.

La regresión PLS es pues un método estadístico multivariante de aplicación actualmente generalizada. Combina y generaliza conceptos de Análisis de Componentes Principales y de análisis de Regresión Lineal Múltiple. Es particularmente útil cuando se desea predecir un conjunto de variables dependientes (**Y**) desde un conjunto (relativamente grande y posiblemente correlacionadas) de variables predictoras (**X**). De este modo es posible maximizar no solo la varianza de las **X** del sistema, sino también la varianza de las **Y**.

Cuando **Y** es un vector y **X** una matriz a rango completo, este objetivo podría ser cumplido usando Regresión Múltiple Ordinaria. Cuando el número de estimadores es grande comparado con el número de observaciones, lo más probable es que **X** sea singular y la aproximación de regresión ya no sea factible. Tal como se comentó en un párrafo anterior, otro problema a enfrentar es la multicolinealidad, que generalmente se soluciona eliminando las variables que la causan o transformándolas. Esta solución resulta efectiva cuando no se pretende que el set de variables X permanezca en el modelo por necesidades de explicación y predicción.

Muchas aproximaciones han sido desarrolladas para enfrentar este problema:

• una aproximación es eliminar algunos estimadores (por ejemplo, usando métodos de aproximación por pasos),

• otra, llamada regresión de componente principal, es ejecutar un Análisis de Componentes Principales (ACP) de la matriz **X** y entonces usar los componentes principales de **X** como una regresión en **Y**. La ortogonalidad de las componentes principales elimina el problema de la multicolinealidad. Pero, el problema es elegir un subconjunto *óptimo* de estimadores residuales. Una posible estrategia es reservar solamente algunas de las primeras componentes. Pero ellas se seleccionan para explicar **X** más que **Y**. Por consiguiente, no hay ninguna garantía de que las componentes principales, que "explican" **X** sean relevantes para **Y**.

Por el contrario, la regresión PLS encuentra componentes de **X** que también son relevantes para **Y**. La regresión PLS reemplaza el espacio inicial de numerosas variables explicativas por un nuevo espacio de menor dimensionalidad definido por un pequeño número de variables, llamadas "factores" o "variables latentes", que son construidas una después de la otra, consecutivamente de forma iterativa. Estos factores, serán las nuevas variables explicativas de un modelo de regresión lineal clásica. Específicamente, busca un conjunto de componentes (*denominados vectores latentes*) que ejecutan una descomposición simultánea de **X** y de **Y** con la restricción de que estos componentes explican tanto como sea posible la *covarianza* entre **X** e **Y**. Este paso generaliza ACP y es seguido por un paso de regresión donde la descomposición de **X** se usa para predecir **Y**.

También puede ocurrir que, además de lo expuesto, el problema requiera de relaciones múltiples y cruzadas, y que todas ellas se den simultáneamente, o que existan variables que no se puedan medir directamente (no observables) y que sean necesarias para desarrollar la teoría.

Sintetizado algebraicamente:

• Los Componentes Principales son los vectores propios de la matriz $\boldsymbol{X'X}$, siendo $\boldsymbol{X'}$ la matriz transpuesta de $\boldsymbol{X}$.

• Entonces los factores de la regresión PLS son los vectores propios de la matriz $\boldsymbol{Y'XX'Y}$ (siendo $\boldsymbol{Y'}$ la transpuesta de $\boldsymbol{Y}$), donde las variables a explicar y las variables explicativas son tomadas en cuenta simultáneamente.

A continuación se exponen algunos ejemplos en los cuales resultaría adecuado aplicar la regresión PLS para, primero describir, y posteriormente predecir el comportamiento de las variables que interesen:

• En estudios relacionados con la contaminación del medio ambiente donde las emisiones de gases contribuyen de manera importante a su deterioro, se pretende explicar en qué medida inciden sobre el índice de octano (Y) de las naftas, cada una de las esencias (X) que la componen. De ser posible determinar también en que proporción debería actuar cada una para encontrar el óptimo octano, para consecuentemente provocar el menor deterioro del medio ambiente[4].

• Una vinoteca desea conocer el éxito comercial de 5 nuevos vinos elaborados por su principal bodega proveedora. Particularmente quisiera saber, de acuerdo a determinadas ca-

[3] Ejemplo cuya solución se ofrece en el presente trabajo.

racterísticas (X) de interés, tales como el contenido de azúcar, de alcohol, la acidez de cada vino y su precio, se adaptan a los postres, carnes y al paladar de los clientes, características(Y), evaluadas por un grupo de expertos[5].

• En un estudio realizado con alumnos del último año de la escuela secundaria, se desea conocer si las calificaciones en asignaturas de naturaleza cuantitativa (como Matemática, Física y Contabilidad) se correlacionan o no con las calificaciones obtenidas en asignaturas de naturaleza no cuantitativa (como Lengua, Literatura e Historia) y, de ser posible, conocidas las calificaciones de unas poder predecir qué podría esperarse en las otras. Los docentes responsables del ensayo opinan que los alumnos que tienen buen desempeño en las materias de naturaleza cuantitativa lo tendrían también en las materias no cuantitativas[6].

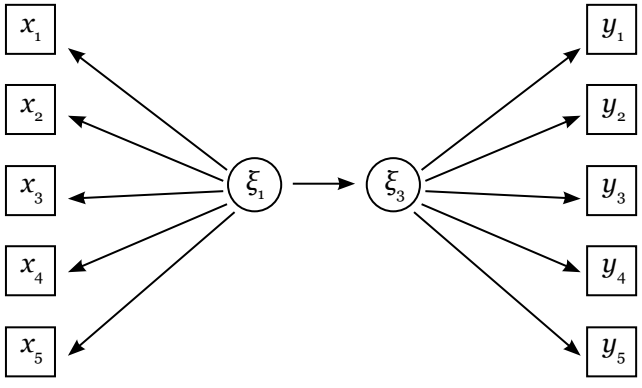
En general se tiene una matriz **X** con **p** variables predictoras $x_1, x_2,..., x_p$ y otra **Y** con **q** variables a explicar $y_1, y_2,....., y_q$, escritas del modo siguiente:

$$X = \{x_1, x_2,...,x_j,...x_p\} \qquad Y = \{y_1, y_2,...,y_j,...y_p\}$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \qquad Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{bmatrix}$$

Gráficamente, la situación podría resumirse mediante el siguiente esquema:

Figura 1: Diagrama de un Modelo General



Fuente: Modelo mediante la notación Tenenhaus. (1993)

[5] Ejemplo cuya solución se encuentra en el presente trabajo

[6] La solución realizada con el software SMART PLS brinda una aproximación del PLS. Será objeto de una próxima publicación.

En la Figura 1, las ξ representan a variables no observables, a "medir" a través de sus indicadores x_p o y_q, respectivamente.

Sobre escalas de medida y distribución de los datos

Como en todo modelo, para lograr la estimación pretendida se requiere la explicitación de supuestos. PLS presenta beneficios al respecto, dado que:

• No implica ningún modelo estadístico y por lo tanto evita la necesidad de realizar suposiciones con respecto a las escalas de medida.

• Por consiguiente, las variables pueden estar medidas por diversos niveles de medida (si las escalas son numéricas se utiliza ACP, en tanto que si las escalas son categóricas u ordinales va implícito el AFCM).

• No requiere distribución conocida de los datos.

3.1 LA REGRESIÓN PLS UNIVARIADA (PLS1)

A efectos de comprender la filosofía PLS, se expone primeramente, el caso en que hay una sola variable a explicar *Y*, metodología que se la designa comúnmente como PLS1 (Tenenhaus, 1998), para luego generalizarlo al caso de más de una variable a explicar, que se designa como PLS2.

Existen numerosas versiones del algoritmo de regresión PLS1 que, si bien difieren en cuanto al nivel de normalización y necesidad de cálculos intermedios, convergen absolutamente a la misma regresión. Sea ***Y*** un vector (*n x 1*) y **X** una matriz de orden (*n x p*). **Y** puede ser visto como una transformación de las variables predictoras ***X***, considerando su relación con el vector de respuestas ***Y***, obteniéndose como resultado una matriz de componentes o variables latentes no correlacionadas $\boldsymbol{T}, / \boldsymbol{T} = (\boldsymbol{t}_1; \boldsymbol{t}_2;...;\boldsymbol{t}_p)$ deben ser de orden $n \times p$. Es de notar que esto contrasta con el ACP, en el cual las componentes son obtenidas utilizando solamente la matriz de predictoras ***X***. El número de variables latentes $t_1; t_2;...;t_k$ a retener, es determinado generalmente por el método de validación cruzada, designado generalmente PRESS (Prediction Sum of Squares).

La ecuación de regresión estimada toma la siguiente forma:

$$\hat{y} = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + ... + \beta_k t_k$$

En este caso se busca realizar una regresión de una variable a explicar, **Y**, sobre **p** variables explicativas **X**: $x_1, x_2,..., x_p$, tal que la matriz **X** y el vector **Y** adoptan la siguiente estructura:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

Como se especificó en párrafos anteriores, las variables **X** pueden estar altamente correlacionadas y al mismo tiempo puede haber más variables que observaciones, pero por otro lado los coeficientes de regresión deben ser interpretables (en el sentido de la práctica), o sea se tiene en cuenta el hecho de buscar en que medida x_j contribuye a la construcción de la variable y con la ayuda del coeficiente de regresión. En los casos en que éste tiene un signo opuesto al correspondiente coeficiente de correlación lineal (x_j, y) presenta dificultades de interpretación. Se pretende lograr el objetivo de una regresión interpretable, para lo cual se aplica el algoritmo que se presenta a continuación:

3.1.1 ALGORITMO DE LA REGRESIÓN PLS1

Una vez introducidos los datos, los pasos a seguir se desarrollan de acuerdo a los pasos detallados a continuación:

Paso 1: Construcción de la primera componente t_1

La primer componente t_1 se define como

$$t_1 = w_{11}x_1 + w_{12}x_2 + \ldots\ldots\ldots + w_{1p}x_p$$

en la cual cada w_{1j} se obtiene mediante la expresión general $w_{1j} = \dfrac{Cov(x_j; y)}{\sqrt{\sum_{j=1}^{p} Cov^2(x_j; y)}}$

En particular

$$w_{11} = \frac{Cov(x_1; y)}{\sqrt{\sum_{j=1}^{p} Cov^2(x_j; y)}} \quad ; \quad w_{12} = \frac{Cov(x_2; y)}{\sqrt{\sum_{j=1}^{p} Cov^2(x_j; y)}} \quad \ldots \quad w_{1p} = \frac{Cov(x_p; y)}{\sqrt{\sum_{j=1}^{p} Cov^2(x_j; y)}}$$

Paso 2: Se efectúa una regresión MCO de y sobre t_1, es decir

$Y_y = f(t_1)$ donde t_1 es de la forma $t_1 = w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p$

En general el modelo de regresión teórico se puede escribir como la regresión

$Y_y = c_1 t_1 + y_1$, donde c_1: *coeficiente de regresión;* y_1*: vector de residuos*

y c_1 se calcula como:

$$c_1 = \frac{Cov(y; t_1)}{\|t_1\|^2}$$

Para el cálculo de los **residuos** y_1, se efectúan previamente las regresiones simples de los x_j sobre t_1 y se obtienen las rectas de predicción estimadas:

$x_j = f(t_1)$ *para* $j = 1,2,\ldots, p$

tal que, cada recta estimada responde a la expresión $\hat{x_j} = \hat{\alpha_j}\, t_1$

Las estimaciones de los coeficientes de regresión se calculan del siguiente modo:

$$\hat{\alpha}_j = \frac{Cov(x_j; t_1)}{\|t_1\|^2} \qquad \text{para } j = 1,2,\ldots, p$$

Una vez obtenidas las estimaciones anteriores, se está en condiciones de calcular los residuos asociados a las rectas de regresión mediante una simple sustracción:

$y_{1,j} = x_j - \hat{x}_j$ donde x_j = *valor observado*

$\hat{x}_j$ = *valor estimado*

Por lo cual una primera ecuación de regresión adopta la forma

$y = c_1 w_{11} x_1 + c_1 w_{12} x_2 + \ldots + c_1 w_{1p} x_p + y_1$

donde los coeficientes son más fáciles de interpretar.

Si el poder explicativo de esta regresión es muy débil ó poco confiable, se busca construir una segunda componente t_2, combinación lineal de las x_j, no correlacionada a t_1, y que explique bien los residuos y_1 , mediante:

$$t_2 = w_{21} x_{11} + w_{22} x_{12} + w_{23} x_{13} + \ldots + w_{2p} x_{1p}$$

donde

$$w_{2j} = \frac{Cov(x_{1j}; y_1)}{\sqrt{\sum_{j=1}^{p} Cov^2(x_{1j}; y_1)}} \text{ ,tal que } w_{21} = \frac{Cov(x_{11}; y_1)}{\sqrt{\sum_{j=1}^{p} Cov^2(x_{1j}; y_1)}} \; ; \quad w_{22} = \frac{Cov(x_{12}; y_1)}{\sqrt{\sum_{j=1}^{p} Cov^2(x_{1j}; y_1)}}$$

Nótese que los w_{2j} se construyen utilizando la variabilidad residual, por lo cual t_2, es una componente principal calculada sobre los **residuos de las x_{ij}**.

Se efectúa después una regresión empleando MCO de y sobre t_1 y t_2, condición que le adjudica el nombre de **Mínimo Cuadrado Parcial**.

$y = f(t_1, t_2) \Rightarrow y = c_1 t_1 + c_2 t_2 + y_2$

Se expresan t_1 y t_2 en función de las variables x_j, la ecuación de regresión se puede escribir en función de esas variables. Donde una segunda ecuación de regresión puede ser más precisa que la primera.

$y = c_1 w_{11} x_1 + c_1 w_{12} x_2 + \ldots + c_1 w_{1p} x_p + c_2 w_{21} x_{11} + c_2 w_{22} x_{12} + + c_2 w_{23} x_{13} + \ldots + c_2 w_{2p} x_{1p} + y_2$

Este procedimiento iterativo puede ser continuado utilizando de igual modo los residuos $y_2, x_{21}, x_{22}, \ldots, x_{2p}$ de las regresiones de $y, x_1, x_2, \ldots, x_p$ sobre t_1 y t_2.

El número de componentes a retener es habitualmente determinado por validación cruzada, tal como, en su forma general, se especificó en Sec. 3.1.

El número de componentes t_1, t_2, ..., t_h a retener es habitualmente determinado por *validación cruzada*, tal como, en su forma general, se especificó en Sec. 3.1.

Por cada valor, se calculan las predicciones $\hat{y}_{hi}$ de y_i con la ayuda del modelo a h componentes, (calculado utilizando todas las observaciones) y posteriormente se realizan las mismas cuentas, las predicciones $\hat{y}_{h(-i)}$, pero, sin utilizar la observación i, es decir la correspondiente al individuo i, indicado como $h(-i)$. Una vez obtenidas las predicciones, se utilizan los criterios RSS_h (Residual Sum of Squares) y $PRESS_h$ (Prediction Error Sum of Squares) definidos por

$$RSS_h = \sum (y_i - \hat{y}_{hi})^2$$

y $$PRESS_h = \sum (y_i - \hat{y}_{h(-i)})^2$$

De acuerdo al algoritmo, la componente t_h es conservada si

$$\sqrt{PRESS_h} \leq 0{,}095 \sqrt{RSS_{h-1}}$$

3.1.2 UN EJEMPLO DIDÁCTICO ILUSTRATIVO

El ejemplo a analizar contiene datos extraídos de una investigación de Cornell (1990), posteriormente vista por Kettaneh & Wold (1992) y retomada por Tenenhaus (1998) donde el objetivo es explicar en qué medida inciden sobre el índice de octano (Y) de las naftas, cada una de las 7 componentes(X) descriptas a continuación. De ser posible determinar en que proporción debería actuar cada una para encontrar el óptimo octano. Para ello se tomó una muestra aleatoria de 12 mezclas con las 7 componentes en las proporciones que se detallan en la Tabla 1. Este problema se originó en estudios relacionados con la contaminación del medio ambiente donde las emisiones de gases contribuyen de manera importante a su deterioro.

Tabla 1: Descripción de las variables

Componentes	Valores
Destilación directa	$0 \leq x_1 \leq 0{,}21$
Reformat	$0 \leq x_2 \leq 0{,}62$
Nafta de craqueado térmico	$0 \leq x_3 \leq 0{,}12$
Nafta de craqueado catalítico	$0 \leq x_4 \leq 0{,}62$
Refinería	$0 \leq x_5 \leq 0{,}12$
Al kylat	$0 \leq x_6 \leq 0{,}74$
Esencia Natural	$0 \leq x_7 \leq 0{,}08$

Fuente: Tenenhaus, (1988)

Las proporciones de componentes correspondientes a cada uno de las 12 muestras están distribuidas, según la Tabla 2. La suma de proporciones x_j para cada mezcla es la unidad.

Tabla 2: Datos de Cornell

Nº	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	0,00	0,23	0,00	0,00	0,00	0,74	0,03	98,7
2	0,00	0,10	0,00	0,00	0,12	0,74	0,04	97,8
3	0,00	0,00	0,00	0,10	0,12	0,74	0,04	96,6
4	0,00	0,49	0,00	0,00	0,12	0,37	0,02	92,0
5	0,00	0,00	0,00	0,62	0,12	0,18	0,08	86,6
6	0,00	0,62	0,00	0,00	0,00	0,37	0,01	91,2
7	0,17	0,27	0,10	0,38	0,00	0,00	0,08	81,9
8	0,17	0,19	0,10	0,38	0,02	0,06	0,08	83,1
9	0,17	0,21	0,10	0,38	0,00	0,06	0,08	82,4
10	0,17	0,15	0,10	0,38	0,02	0,10	0,08	83,2
11	0,21	0,36	0,12	0,25	0,00	0,00	0,06	81,4
12	0,00	0,00	0,00	0,55	0,00	0,37	0,08	88,1

De acuerdo a los objetivos planteados, la solución podría encontrarse mediante una regresión múltiple, con la aplicación de MCO.

A tal efecto, se realiza primeramente una RLM, con sus correspondientes limitaciones para luego aplicar PLS1 y realizar las comparaciones aludidas en el trabajo.

Previo a la estimación del modelo RLM, se presenta en la Tabla 3, la matriz de correlaciones correspondiente a la totalidad de las variables objeto de análisis, con el fin de conocer el sentido y la intensidad de las relaciones lineales.

Tabla 3: Matriz de correlaciones

Nº	X2	X3	X4	X5	X6	X7	*Y*
X1	,114	1,00(**)	,371	-,548	-,805(**)	,603(*)	-,837(**)
X2		,111	-,533	-,297	-,198	-,584(*)	-,079
X3			,374	-,548	-,805(**)	,607(*)	-,838(**)
X4				-,211	-,646(*)	,916(**)	-,707(*)
X5					,463	-,274	,494
X6						-,656(*)	,985(**)
X7							-,741(**)

** La correlación es significativa al nivel 0,01 (bilateral).
* La correlación es significativa al nivel 0,05 (bilateral).

Según la lectura de la matriz, se desprende que la variable X5 está poco correlacionada con el conjunto de las otras variables, con ninguna correlación significativa a los niveles establecidos; para las variables X1, X3 y X7 se presentan problemas de multicolinearidad.

a) La Regresión por Mínimos Cuadrados Ordinarios

Si bien la matriz de correlaciones indica que se presentan problemas para la estimación propuesta, igualmente se ajusta el modelo sobre el total de las variables explicativas, objetivo del problema. Los resultados se resumen la Tabla 4.

Tabla 4: Resumen de Coeficientes Estimados

Modelo	Coeficientes no estandarizados		Valor del estadístico t	Probabilidad
	B	Error Típico	B	
Constante	94,314	56,133	1,680	**,168**
X1	-73,922	276,807	-,267	**,803**
X2	-8,429	56,368	-,150	**,888**
X3	72,447	433,919	,167	**,876**
X4	-16,718	54,379	-,307	**,774**
X5	-6,400	55,439	**-,115**	**,914**
X6	6,152	55,196	,111	**,917**
X7	18,832	121,423	**,155**	**,884**

Fuente: Elaboración propia con SPSS 17
R = 0,996 (coeficiente de correlación múltiple)
F = 110,62(Nivel de significación = 0,000). Test de Fisher, evalúa la significación global de la regresión.

El resumen muestra que ningún coeficiente es significativo y que los signos de algunos coeficientes no coinciden con los de las correlaciones entre las variables *Xj* e Y. Estos resultados se pueden mejorar desde el punto de vista estadístico efectuando una regresión por pasos descendentes. Las variables retenidas, luego de 4 pasos, se resumen en la Tabla 5.

Tabla 4: Resumen de Coeficientes Estimados

Modelo	Coeficientes no estandarizados		Valor del estadístico t	Probabilidad
	B	Error Típico	B	
Constante	100,995	,752	134,268	,000
X1	-39,798	3,005	-13,245	,000
X2	-15,215	1,538	-9,890	,000
X4	-21,949	1,356	-16,186	,000
X5	-12,736	4,877	-2,611	,035

Como puede observarse, en este caso, todos los coeficientes resultan significativos. No obstante, este modelo si bien satisface requerimientos de la estadística, no cumple con las demandas de la química. Desde el punto de vista de ésta, no es aceptable la eliminación de la variable X_6 del modelo dado que es la más correlacionada con *Y*. Además de no ayudar a la obtención de una mezcla que permita tomar una decisión sobre las siete componentes.

Se justifica de tal forma el interés y relevancia que adquiere la regresión PLS1. En síntesis, permite relacionar a la variable Y con el conjunto de las siete variables explicativas a través del empleo del algoritmo descripto en el apartado 3.1.1.

b) Aplicación del Algoritmo PLS1

Paso 1:

En la primera etapa las covarianzas son iguales a las correlaciones, ya que se opera sobre la matriz de datos estandarizados. La componente t_1, es entonces definida como sigue:

$$t_1 = \frac{\left[\sum_{j=1}^{7} Cov(x_j, y)\, x_j\right]}{\sqrt{\sum_{j=1}^{7} Cov^2(x_j, y)}} =$$

$$= \frac{(-0{,}837_{x1} - 0{,}070_{x2} - 0{,}838_{x3} - 0{,}706_{x4} + 0{,}493_{x5} + 0{,}985_{x6} - 0{,}741_{x7})}{1{,}916} =$$

$$= -0{,}437_{x1} - 0{,}037_{x2} - 0{,}437_{x3} - 0{,}368_{x4} + 0{,}257_{x5} + 0{,}514_{x6} - 0{,}386_{x7}$$

La regresión de y sobre t_1, proporciona la siguiente ecuación de regresión:

$\hat{y} = c_1 t_1 = 0{,}4820\, t_1 =$

$= -0{,}2106_{x1} - 0{,}017_{x2} - 0{,}211_{x3} - 0{,}177_{x4} + 0{,}1242_{x5} + 0{,}247_{x6} - 0{,}186_{x7}$

La bondad de la misma se mide mediante: R = 0,961, F = 133

Esta primera ecuación de regresión es mejor que la obtenida mediante regresión múltiple dado que contiene a todas las variables x_j , y presenta la ventaja de ser perfectamente coherente con la investigación, pues los signos de los coeficientes estimados coinciden con los correspondientes a los de sus correlaciones (Tabla 3). Si bien con la primera componente ya se logró un modelo interpretable y con un alto R, igualmente se calculan otras dos componentes, para mostrar el proceso metodológico y también aumentar, si es posible, el R (aunque siempre es recomendable al respecto, recordar el principio de Parsimonia).

Paso 2:

En el caso que el poder explicativo de la primera ecuación sea demasiado débil, lo que es apreciado en función del test F y R calculados, se puede mejorar ligeramente la regresión buscando la segunda componente t_2. La componente t_2, combinación lineal de las x_j, no está correlacionada con t_1, y explica bien los residuos x_{1j} de las regresiones de las variables x_j sobre la componente t_1. Para ello se calculan los *residuos* $y_1, x_{11},..., x_{17}$ de las regresiones de $y, x_1 ,..., x_7$ sobre t_1, para luego calcular la componente t_1, como sigue:

$$t_2 = \frac{\left[\sum_{j=1}^{7} Cov\,(x_{1j}, y_1)\, x_{ij}\right]}{\sqrt{\sum_{j=1}^{7} Cov^2\,(x_{1j}, y_1)}} =$$

es decir,

$t_2 = w_{21}\, x_{11} + w_{22}\, x_{12} + ... + w_{2p}\, x_{1p}$ donde x_{1j} e y_1 son residuos.

Para el ejemplo:

$$t_2 = \frac{(-0{,}0317x_{11} - 0{,}1315x_{12} + 0{,}0326x_{13} - 0{,}0245x_{14} - 0{,}0701x_{15} + 0{,}0999x_{16} - 0{,}0493x_{17})}{0{,}193}$$

Reemplazando a los x_{1j} en función de los x_j y de t_1, se obtiene la segunda componente:

$$t_2 = 0{,}1203x_1 - 0{,}6847x_2 + 0{,}1248x_3 - 0{,}1641x_4 - 0{,}3370x_5 + 0{,}5692x_6 + 0{,}2164x_7$$

Entonces, la regresión de $y = f(t_1, t_2)$, proporciona un R = 0,9881; F= 20,6. La ecuación correspondiente es

$y = c_1\, t_1 = 0{,}4820\, t_1 + 0{,}273t_2 =$

$= -0{,}177x_1 - 0{,}204x_2 - 0{,}176x_3 - 0{,}222x_4 + 0{,}032x_5 + 0{,}403x_6 - 0{,}127x_7$

En la cual $c_2 = \frac{Cov\left[y_1, t_2\right]}{\|t_2\|^2} = \frac{\sqrt{n-1}}{\|t_2\|}.r_{y1,\,t2}$

Paso 3:

De ser necesaria una tercera componente t_3, esta se calcula de modo similar a t_2, es decir, en función de los residuos: y_2, x_{21}, ..., x_{27} de las regresiones de y, x_1, x_2,..., x_7 sobre t_1, t_2.

Entonces la componente t_3 puede escribirse como:

$$t_3 = \frac{\left[\sum_{j=1}^{7} Cov\,(x_{2j}, y_2)\, x_{2j}\right]}{\sqrt{\sum_{j=1}^{7} Cov^2\,(x_{2j}, y_2)}}$$

realizando los pasos correspondientes y expresando t_1 y t_2 en función de las variables x_j, se obtiene:

$$t_3 = 0{,}375x_1 - 0{,}037x_2 + 0{,}380x_3 - 0{,}684x_4 - 0{,}685x_5 + 0{,}515x_6 - 0{,}155x_7$$

La ecuación de regresión a estimar es función de tres componentes: $\hat{y} = f(t_1, t_2, t_3)$, una función que se logra mediante

$$\hat{y} = c_1 t_1 + c_2 t_2 + c_3 t_3 = 0{,}482 t_1 + 0{,}273 t_2 + 0{,}103 t_3 =$$
$$= - 0{,}139 x_1 - 0{,}208 x_2 - 0{,}137 x_3 - 0{,}293 x_4 - 0{,}038 x_5 - 0{,}456 x_6 - 0{,}143 x_7 \ (*)$$

A excepción de la variable x_5, donde el coeficiente es despreciable pero contribuye igualmente a la explicación de $\hat{y}$, la ecuación de regresión conserva su coherencia y la correlación múltiple es máxima (R = 0.996).

Para la lectura de (*) cada coeficiente de regresión mide la contribución de la variable x_j a la construcción de la variable y.

Siguiendo el criterio de retener aquellas h primeras componentes que verifiquen que $Q^2_h > 0.0975$, la validación cruzada, según el índice de Stone-Geiger[7] y de acuerdo al número de componentes principales retenidas, se obtienen los siguientes resultados[8], según se hayan retenido una, dos o las tres componentes:

$Q^2_1 = 0.987$; $Q^2_2 = 0.202$ y $Q^2_3 = 0.272$

Estos valores indican que en cualquiera de los tres casos se satisface la condición PRESS, pero a efectos de trabajar con el más alto valor de R, se hace la predicción con la ecuación a tres componentes.

Predicción

Si se supone que se trata de fabricar una mezcla de componentes que conduzca al valor máximo de octano (o sea Y), considerando la ecuación estimada a tres componentes (*) previa reconstitución de los datos activos (variables originales, dado que para la estimación se operó con las variables estandarizadas), resulta la siguiente expresión:

$$\hat{y} = 94{,}314 - 9.828 x_1 - 6.96 x_2 - 16.67 x_3 - 8.422 x_4 - 4.389 x_5 + 10.16 x_6 - 33.53 x_7$$

Para maximizar el índice y se puede construir una mezcla donde prevalezcan los valores de las componentes con coeficientes más elevados, respetando la suma igual a la unidad.

[7] Q^2_h, Índice de Stone Geiser mide el aporte marginal de cada componente PLS *th* al poder predictivo del modelo. El aporte de *th* es significativo si $Q^2_h \geq (1 - 0.95) = 0.0975$. Límite que corresponde al creador de la teoría PLS, Wold, H.

[8] A efectos de los cálculos numéricos dirigirse a los autores del trabajo, o consultar Valencia Delfa & Díaz Llanos, (2003), pp. 48 - 77

3.2 LA REGRESIÓN PLS2

Se entiende como regresión PLS2 (Tenenhaus, 1998) la extensión de la regresión PLS1 a la situación en la que se tenga un conjunto de variables a explicar Y =$\{y_1, y_2,..., y_q\}$, que se trata de relacionar con otro conjunto de variables explicativas ó predictoras X =$\{x_1, x_2,..., x_p\}$. Por tanto, la regresión PLS2 consiste en efectuar un análisis de componentes principales de un conjunto de variables X, bajo la condición de que estas componentes principales brinden también la mejor explicación posible respecto del conjunto de variables Y. En este caso es posible predecir las variables y_k a partir de las x_j diferenciando mejor lo que es común a los datos del problema de aquello que es más específico respecto de Y.

Este procedimiento se entiende más fácilmente desde el punto de vista geométrico, donde se pueden ver las matrices X e Y como nubes de puntos en dos espacios, el espacio X con p-ejes ($\Re^p$) y el espacio Y con q-ejes ($\Re^q$), donde p y q son las columnas de X e Y respectivamente (número de variables). En este contexto la modelización PLS2 consiste en proyecciones simultáneas de ambos espacios sobre subespacios de menor dimensión. Las coordenadas de los puntos en estos subespacios constituyen los elementos de las matrices T y U (matrices de componentes) que se obtienen del análisis.

Por lo que se consiguen dos objetivos primordiales:

Maximizar la correlación entre los conjuntos de variables X e Y.

Aproximar a través del citado subespacio lo mejor que se pueda a los espacios generados por las variables X e Y (es decir, la información que poseen).

Este método es efectivo cuando los conjuntos de variables antes citados son muy numerosos, el conjunto de variables X = $\{x_1, x_2,..., x_p\}$ se encuentra muy correlacionado entre sí y a su vez también correlacionado con el conjunto Y = $\{y_1, y_2,..., y_q\}$.

3.2.1 UN EJEMPLO SENCILLO DE APLICACIÓN DE PLS2

Una vinoteca desea conocer el éxito comercial de 5 nuevos vinos elaborados por su principal bodega proveedora. Para ello se plantea predecir la evaluación subjetiva de un conjunto de 5 vinos[9], donde las variables dependientes a predecir para cada vino son su calidad o sabor, adaptación a las carnes o a los postres, según fueron evaluados por un panel de expertos. Como variables características o explicativas se consideraron el precio, el contenido de azúcar, de alcohol y la acidez de cada vino.

[9] Ejemplo extraído de Lewis-Beck et al.(2003)

Los datos se encuentran en las Tablas 6 y 7, respectivamente:

Tabla 6: La matriz Y de variables dependientes

Vino	Calidad	Carnes	Postres
1	14	7	8
2	10	7	6
3	8	5	5
4	2	4	7
5	6	2	4

Tabla 7: La matriz X de predictores

Vino	Precio	Azúcar	Alcohol	Acidez
1	14	7	8	7
2	10	7	6	7
3	8	5	5	5
4	2	4	7	3
5	6	2	4	3

Tabla 8: La matriz de resultados

	Calidad	Carnes	Postres
Precio	**-0,2662**	-0.2498	0.0121
Azúcar	0.0616	0.3197	**0.7900**
Alcohol	**0.2969**	0.3679	0.2568
Acidez	0.3011	0.3699	0.2506

La tabla precedente presenta las estimaciones mediante procedimiento PLS2, con el empleo de sólo dos vectores latentes (componentes principales), dado que la variabilidad explicada de **Y** es del 85%, en tanto que la de **X** es del 98%, porcentajes considerados suficientes por quienes realizaron el trabajo, además de que siempre es preferible aplicar el principio de parsimonia.

Los resultados exhibidos en las columnas $\hat{y}_1$ (Calidad), $\hat{y}_2$ (Carnes) e $\hat{y}_3$ (Postres) corresponden a los "pesos" (coeficientes de regresión estandarizados) con los cuales cada variable explicativa contribuye a la explicación de la **Y** y, en consecuencia, a su predicción. Es así como el azúcar es el principal responsable en la elección de un vino de postre y que el precio se asocia negativamente a la calidad percibida del vino, mientras que el alcohol está relacionado positivamente con la calidad.

3.3 UN ALGORITMO DE APROXIMACIÓN AL MODELO PLS

En el desarrollo del presente trabajo, se mostró la estimación PLS sólo para problemas en las cuales la totalidad de las variables empleadas son observables y expresadas numéricamente (variables cuantitativas), características que explican la utilización del ACP para obtener los vectores latentes.

También se puede realizar una estimación PLS cuando las variables explicativas son de carácter cualitativo, en cuyo caso los vectores latentes necesitan del empleo del Análisis Factorial de Correspondencias Múltiples (AFCM) (Cazes, 1997).

En los últimos años, surge una aproximación a PLS que permite realizar estimaciones donde intervienen variables no observables en el conjunto de variables explicativas como también en el conjunto de variables a explicar. Es una presentación de la aproximación de Cuadrados Mínimos Parciales a la Modelización de Ecuaciones Estructurales (ó Modelización PLS Path). Esta aproximación se compara con la estimación de la Modelización de Ecuaciones Estructurales a través de la máxima verosimilitud, se puede usar para analizar tablas múltiples tanto así como para relacionarlas a métodos de análisis de datos más clásicos en este campo.

En este caso, el algoritmo PLS es una secuencia de regresiones en términos de los vectores peso. Los vectores peso obtenidos en una convergencia, satisfacen las ecuaciones de puntos fijos (para un análisis general de dichas ecuaciones siguiendo los casos de convergencia, ver Henseler, et al., (2010)). El algoritmo PLS básico, tal como sugiere Lohmöller (1989), incluye las siguientes tres etapas:

Etapa 1: *Estimación iterativa de valores de variables latentes*, que consiste en un procedimiento iterativo de cuatro pasos que se repite hasta obtener la convergencia.

Etapa 2: *Estimación de los pesos/cargas externas y coeficientes de dependencia*.

Etapa 3: *Estimación de parámetros de posición*.

Este conjunto relativamente sencillo de regresiones simples y múltiples puede ser extendido a los modelos causales complejos, a medida que el algoritmo PLS toma segmentos de modelos complejos y aplica el mismo proceso hasta que el modelo completo converge. De esta forma, en un momento determinado, el procedimiento iterativo está trabajando con un constructo y un conjunto de medidas o variables observables relacionadas con este constructo, o con constructos adyacentes en el modelo.

Es esta segmentación de modelos complejos lo que permite que PLS opere con muestras pequeñas cuando se requiere el uso de componentes principales, como ocurre en este caso.

4. CONCLUSIÓN

La regresión PLS permite una mejor estimación de los estadísticos empleados por el investigador en los problemas de modelización. En efecto, el investigador pretende conservar en el modelo todas las variables importantes, es decir, todas aquellas que resultaron coherentes al ser obtenidas en la ecuación de regresión.

Cuando se emplea una regresión múltiple, si existe colinealidad y/o una cantidad importante de variables explicativas en relación al número de observaciones, la solución más corriente consiste en excluir las variables explicativas mediante los métodos *paso a paso*. La regresión PLS permite, en tales casos, como ha sido demostrado, conservar todas las variables explicativas obtenidas en una ecuación de regresión coherente, lo que implica una importante contribución en procesos de investigación aplicada.

En otro trabajo se podrá profundizar esta línea de investigación abordando la estimación mediante el empleo del algoritmo de aproximación PLS con la intervención de variables no observables.

REFERENCIAS BIBLIOGRÁFICAS

Albano, C., Blomquist, G., Dunn III,W. & Edlund, U. (1983) *Régression PLS et Applications* - Revue de Statistique Appliquée, Vol 43, Nº1, p.p 65-89.

Cazes, P. (1997) *Adaptation de la Régression PLS au Cas de la Régression après Analyse des Correspondances Multiples* – Revue de Statistique Appliquée, Vol. 45 (2).

Cornell, J. (1990) *Experiments with Mixtures: Designs, Models and the Analysis of Mixture data* – John Wiley & Sons, New York.

Garthwaite, P. (1994) *Statistical Inference* – Oxford University Press.

Gujarati, D. (2004) *Introducción a la Econometría* – Ed. McGraw-Hill, México.

Hellands, I. (1988) *On the Structure of Partial Least Squares Regression* Communication in Statistics, Simulación and Computation, Vol. 17, Nº2.

Henseler, J., Ringle, C. & Sarstedt, M. (2010) *Using partial least squares path modeling in international advertising research: Basic concepts and recent issues* – Okazaki (Ed.), Handbook of Research in International Advertising.

Hotelling, H. (1933) *Analysis of a complex of statistical variables into principal component* – Journal of Educational Psychology, Vol 24(7), Oct 1933, 498-520.

Höskuldsson, A.(1988) *PLS Regression Methods* – Journal of Chemometrics, Vol.2.

Kettaneh & Wold, H. (1992) *Analysis of mixture data with Partial Least Squares* – Chemometrics and Intelligent Laboratory Systems, Vol. 14, pp. 57-69.

Langrand, C., Pinzón, L.M. (2009) *Análisis de Datos, Métodos y Ejemplos* – Ed. Escuela Colombiana de Ingeniería, Colombia.

Lewis-Beck, M. et al. (2003) *Partial Least Squares (PLS) Regression* – Encyclopedia of Social Sciences Research Methods, Thousand Oaks (CA): Sage.

Lohmöller, J. (1989) *Latent Variable Path Modeling with Partial Least* Squares – Heidelberg, Germany. Physicaverlag.

Martens, H. & Naes, T. (1989) *Multivariate Calibration* – John Wiley & Sons, New York.

Pearson, K. (1901) *On Lines and Planes of Closed Fit to System of Point in Space* – The Philosophical Magazine, Series 6-2(11), pp. 559-572.

Plá, L. (1986) *Análisis Multivariado: Método de Componentes Principales* – Secretaría General de la Organización de los Estados Americanos, Washington, D. C.

Pulido, A., López, L. (1999) *Predicción y Simulación Aplicada a la Economía y Gestión de Empresas* – Editorial Pirámide, Madrid.

Pulido, A. (1987) *Predicción y Simulación Aplicada a la Economía* – Editorial Pirámide, Madrid.

Tenenhaus, M. (1993) *La Régression PLS Généralisée* – Cahier de Recherche Nº 472, Groupe HEC, Jouy-en-Josas.

Tenenhaus, M. (1998) *La Régression PLS.Théorie et Pratique* – Editions Technip, París.

Valencia Delfa, J., Díaz Llanos, J. & Calleja-Sáinz (2003) *PLS en las Ciencias Experimentales* – Ed. Editorial Complutense, Madrid.

Wold, H. (1975) *Modelling in Complex Situations with Soft Information*. Third World Congress of Econometric Society, August 21-26, Toronto, Canadá.

Wold, H. (1985) *Partial Leas Squares, en Encyclopedia of Statistical Sciences, vol 6*. Editorial John Wiley & Sons, New York.

Wold, S., Martens, H. & Wold, H. (1983) *The Multivariate Calibration Problem in Linear Regression. The Partial Least Squares (PLS)* - In Proc.Conf. Matrix Pencils Editorial Ruhe, A. & Kägstrom, B.

CURRICULUM VITAE

Ana Maria Legato

Magister en Administración de Negocios (UNICEN, Argentina-California State University, Los Ángeles, EEUU).
Profesor Titular Dedicación Exclusiva, Facultad de Ciencias Económicas-UNICEN.
Profesor Invitado para cursos de Doctorado en la Facultad de Económicas y Empresariales, Universidad de Castilla La Mancha – España, sobre Técnicas Avanzadas de Investigación: Análisis de Datos Multivariados.
Profesor del Doctorado en Ciencias de la Gestión, Facultad de Ciencias Económicas-UNLP, sobre Métodos Cuantitativos en Investigación.
Profesor responsable de cursos en la Maestría en Dirección de Empresas-UNLP, sobre Estadística y Probabilidades en los Negocios y Pronósticos en los Negocios.
Autor de Textos y publicaciones sobre la especialidad.
Estudios de Posgrado en el Exterior (Università di Siena, Italia).
Investigador del CEA(Centro de Estudios en Administración), Facultad de Ciencias Económicas-UNICEN.
legato@econ.unicen.edu.ar

Aldo Hernan Alonso

Doctor en Ciencias Económicas (UNLP).
Profesor Titular de Finanzas de Empresa (UNLP, UNICEN).
Profesor Consulto (UNLP).
Profesor de Finanzas Empresariales en Posgrados (Maestrías y/o de especialización en la UBA, UNICE, UNLP, UNdeSL, UNdeER, UCACER, UNNE, UNdelSUR).
Director de la maestría en Dirección de Empresas ofrecida por la UNLP.
Director del Posgrado de Especialización en Gestión de la Empresa Agropecuaria (UNICEN y UNNE).
Autor de Textos y publicaciones sobre la especialidad.

Asesor del CEA (Centro de Estudios en Administración), Convenio UNICEN-UNLP.
Presidente de la Sociedad Argentina de Docentes en Administración Financiera (SADAF), períodos 1995/97 y 1997/99.
Estudios de Posgrado en el Exterior (Kansas University, USA).
Profesor visitante en la Universidad Carlos III (Madrid) y Universidad de Castilla-La Mancha (Toledo y Albacete), España
director@mbaunlp.com.ar
aldohalonso@yahoo.com.ar