# Modelos de Regresión PLS Aplicados a Variables Educativas

## PLS Regression Models Applied to Educational Variables

Carlos Daniel Acosta Medina, Germán Albeiro Castaño Duque, Jaider Albeiro Figueroa Flórez
*Departamento de Matemáticas y Estadística, Universidad Nacional de Colombia, Manizales, Colombia*
Correo-e: cdacostam@unal.edu.co, gacastanod@unal.edu.co, jaiderfig@yahoo.com

***Resumen*— Se realiza un estudio sobre el comportamiento de variables educativas asociadas a indicadores de alta calidad en los programas de pregrado de la Universidad Nacional de Colombia - Sede Manizales, con el objeto de crear modelos de regresión multivariado que permitan proyectar comportamientos futuros y establecer relaciones entre las variables asociadas a los factores estudiantes, docencia, procesos académicos e investigación, y las relacionadas con garantía, reconocimiento y aseguramiento de la calidad. Para la construcción y estructuración de los modelos se utilizan las técnicas PLS y Kernel PLS, haciendo los ajustes y validaciones pertinentes. Se obtienen modelos que contribuyen al mejoramiento de los aspectos predictivo y explicativo conjuntamente, obteniendo información interesante para tomar decisiones en los ámbitos académico y administrativo.**

***Palabras clave*— *Deflactar, Métodos Kernel, Porcentaje de predicción, Regresión en mínimos cuadrados parciales, Validación cruzada.***

***Abstract*— We study the behavior of educational variables associated to high-quality indicators of the undergraduate programs offered by the Universidad Nacional de Colombia – Sede Manizales, with aims at elaborating multivariate regression models that enable us to plan future behavior, and establish relations between the variables associated to the factors students, teaching, academic processes and research, and the ones related to guarantee, recognition and assurance of quality. We use the PLS and Kernel PLS techniques, with the appropriate adjustments and validations. We obtain models that contribute to the improvement of predictive and explicative aspects jointly, obtaining relevant information to be taken into account when making academic or administrative decisions.**

***Key Word*— *Cross-validation, Deflate, Kernel methods, Partial least square regression, Prediction percentage.***

## I. INTRODUCCIÓN

En este trabajo se presentan la construcción, estructuración, implementación, validación y ajustes de los modelos de regresión multivariado lineales y no lineales, aplicando la técnica de mínimos cuadrados parciales PLS (*en inglés Partial Least Squares*) a un estudio sobre el comportamiento de variables educativas asociadas a indicadores de calidad en los programas de pregrado de la Universidad Nacional - Sede Manizales. Se eligen aquellos modelos que expliquen de mejor manera la relación entre las variables estudiantes, docencia, procesos académicos e investigación, y las relacionadas con garantía, reconocimiento y aseguramiento de la calidad educativa en educación superior. Este trabajo surge de la necesidad de priorizar el mejoramiento en los factores de más impacto en calidad, y tomar decisiones adecuadas para el fortalecimiento de los procesos académicos, administrativos y financieros en los programas evaluados.

Entre las ventajas que ofrece el trabajo con la técnica PLS y el uso de los modelos de Regresión PLS (PLSR) [6,7,8,9], encontramos:

- Trabajar con base de datos de las que se desconoce el tipo de distribución probabilística que tienen asociada sus variables.
- Manipular bloques de datos en las que el número de variables es mayor que el número de observaciones.
- Eliminar el problema de multicolinealidad entre las variables explicativas y entre las variables de respuesta, a partir de la construcción de variables latentes ortogonales.
- Las bondades de la técnica PLS para establecer relaciones explicativas entre las variables de entrada y las de respuesta, gracias al criterio de optimización que utiliza (máxima covarianza entre las variables latentes de entrada y las latentes de salida).

En el capítulo II se presentan los conceptos preliminares. En el capítulo III se explica la metodología usada en los procesos de elección de variables, recopilación y adecuación de base de datos, estructuración, ajustes y validación de los modelos. En el capítulo IV se describen los resultados obtenidos en la implementación de los modelos y las discusiones. Finalmente en el capítulo V se presentan las conclusiones del trabajo.

## II. PRELIMINARES

A continuación se exponen los resultados más relevantes que han permitido el desarrollo, validez y avances de la técnica PLS y los modelos PLSR.

Fecha de Recepción: 29 de Junio de 2015
Fecha de Aceptación: 28 de Mayo de 2016

### A. La técnica PLS y el algoritmo NIPALS

Considere $X$ la matriz de datos de tamaño $n \times N$ centrada y rango $a$. La idea básica es descomponer la matriz $X$ en la forma $X = TP^T + E$, donde $T$ está formada por columnas de vectores latentes o componentes ortogonales (*scores*) y es de tamaño $n \times a$, $P$ es la matriz de vectores de peso (*loadings*) de tamaño $N \times a$ y $E$ es la matriz de residuales. Así,
$X = [X_{*1}|X_{*2}|X_{*3}| \dots |X_{*N}] = T_{*1}P_{*1}^T + T_{*2}P_{*2}^T + \cdots + T_{*a}P_{*a}^T$ (2.1), donde $T_{*k}$ y $P_{*k}^T$ son las columnas de las matrices $T$ y $P^T$ respectivamente. $k = 1, \dots, a$.

El primer problema que resuelve PLS es el de multicolinealidad entre las columnas de variables que componen $X$, de modo que cada variable latente se construye de la forma $T_{*k} = Xw$, donde $w$ es un vector adecuado de peso; al deflactar $X$ se garantiza que las ${T_{*k}}'s$ sean ortogonales [5].

De (2.1) y considerando una primera aproximación de $X$ con su primera componente ortogonal $k = 1$, observamos que la columna j-ésima de $X$ tiene la forma $X_{*j} = p_{1j}T_{*1}$, de donde $p_{1j} = {X_{*j}}^T T_{*1}$. Usando una nueva aproximación de $X$ tomando dos componentes principales, obtenemos que la columna $j$-ésima de $X$ tiene la forma $X_{*j} = p_{1j}T_{*1} + p_{2j}T_{*2}$, así $p_{2j} = (X_{*j} - p_{1j}T_{*1})^T T_{*2}$. De modo que para la componente $k$-ésima, la columna $j$-ésima de $X$, tiene la forma: $X_{*j} - p_{1j}T_{*1} = p_{kj}T_{*k}$ de donde

$$p_{kj} = \left(X_{*j} - \sum_{l=1}^{k-1} p_{1j}T_{*l}\right)^T T_{*k} \quad (2.2)$$

Las características de los ${T_{*k}}'s$ y la consecución de los $p_{k*}$ por medio de (2.2) a medida que se consideran nuevas componentes ortogonales, permite construir las matrices $T$ y $P$, obteniendo una aproximación de la matriz original $X$ en la forma $X \approx TP^T$.

El comportamiento descrito en las líneas anteriores constituye la base para la construcción del algoritmo NIPALS (Non-linear Iterative Partial Least Squares) y el desarrollo de los algoritmos PLSR [5].

**Algoritmo NIPALS**

Paso 1. $X_0 = X$ , $X$ centrada o estandarizada.
Paso 2. Para $h = 1, \dots, a$ ($a$ es el rango de $X$).
Paso 3. $t_h$: Inicial.
Paso 4. $p_h = X_{h-1}^T t_h / (t_h^T t_h)$.
Paso 5. Normar $p_h$ a 1.
Paso 6. $t_h = X_{h-1} p_h$.
Paso 7. Deflactar $X, X_h = X_{h-1} - t_h p_h^T$.
Paso 8. Repetir Pasos 3 a 7 hasta convergencia.

### B. Regresión PLS (PLSR)

Considere ahora dos bloques de variables $X$ e $Y$, donde $X$ es la matriz $n \times N$ de variables de entrada o explicativas e $Y$ la matriz de variables de salida o respuesta de tamaño $n \times L$, considere además el problema de multicolinealidad.

El método PLS sugiere construir componentes ortogonales (vectores latentes o vectores scores) en $X$ e $Y$ de la forma $t = Xw$ y $u = Yc$, respectivamente, donde $w$ y $c$ son vectores de peso adecuados de norma 1. Cada vector columna $t$ genera la matriz $T$, de manera análoga cada vector columna $u$ genera la matriz $U$, permitiendo la descomposición de $X$ y $Y$, en la forma [8, 9, 10]

$$X = TP^T + E \quad (2.3)$$
$$Y = UQ^T + F \quad (2.4)$$

donde $T$ y $U$ son matrices de tamaño $n \times a$ cuyas columnas son las componentes ortogonales anteriormente descritas. La matriz $P$ de tamaño $N \times a$ y $Q$ de tamaño $L \times a$ representan las matrices de pesos (*loadings*), y las matrices $E$ y $F$ de tamaño $n \times N$ y $n \times L$, respectivamente, son matrices residuales.

La gran virtud de la técnica PLS es considerar el problema explicativo entre $X$ e $Y$, a partir de las nuevas variables latentes representativas $t = Xw$ y $u = Yc$. La técnica considera un modelo de regresión adecuado, aquel que además de hacer reducción de dimensión garantice una relación explicativa entre estas nuevas variables, por lo que se propone maximizar el cuadrado de la covarianza entre ellas, es decir, se propone resolver el problema [5,12]

$$max_{\|w\|=\|c\|=1}[cov(Xw, Yc)]^2 \quad (2.5)$$

Encontrando los $w$ y $c$ adecuados que formarán las matrices $W$ y $C$, y generando las matrices $T$ y $U$, obtenemos la matriz de coeficientes $B$, de tal modo que el modelo de regresión PLS toma la forma

$$Y = XB + \varepsilon \quad (2.6)$$

### C. Regresión PLS, caso $L = 1$ (PLS1R)

Consideremos el caso univariado ($L = 1$), es decir, una sola variable de salida $Y$. En este caso se busca obtener un modelo de la forma

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_N X_N + \varepsilon \quad (2.7)$$

$X_1, \dots, X_N$ pueden estar altamente correlacionadas.

Se construyen las componentes ortogonales $t = Xw$, realizando en el nuevo espacio de variables regresiones de la forma

$$Y = c_1 t_1 + c_2 t_2 + c_3 t_3 + \cdots + c_a t_a + \delta \quad (2.8)$$

Para resolver (2.5), consideremos $v$ el vector de covarianzas entre $X$ y $Y$, es decir, $v = X^T Y$ con $X$ e $Y$ centradas o estandarizadas; entonces

$$[cov^2(Xw, Y)] = [w^T cov(X, Y)]^2 = [w^T v]^2 = w^T v v^T w \quad (2.9)$$

La función lagrangiana que maximiza (2.9) sujeta a $\|w\| = 1$ es

$$\phi(w, \lambda) = [w^T v v^T w] - \lambda(w^T w - 1) \quad (2.10)$$

Derivando respecto a $w$ e igualando a cero tenemos

$$\frac{\partial \phi}{\partial w} = 2vv^T - 2\lambda w = 0 \quad (2.11)$$

Por tanto,

$$vv^T = \lambda w, \quad (2.12)$$

que es un problema de valores y vectores propios, donde $\lambda$ y $w$ son el autovalor y autovector de $vv^T$ respectivamente.

Premultiplicando (2.12) por $w^T$ e igualmente por $v^T$se deduce de $\lambda$, que el $w$ buscado tiene la forma

$$w = \frac{v}{\|v\|} = \frac{X^T y}{\|X^T y\|} \quad (2.13)$$

que corresponde al vector de covarianzas normalizado [12].

Este resultado se aprovecha en el desarrollo del algoritmo iterativo para PLSR, tomando como base el algoritmo NIPALS en la construcción de las componentes scores. Existen muchas versiones de estos algoritmos, a continuación se presenta una de ellas derivada de la versión SIMCA-P (Software desarrollado por la compañía Umetrics que trabaja básicamente con métodos PCA y PLSR).

**Algoritmo PLS1R**

Se usa en este caso $y$ en vez de $Y$.
Paso 1. $X_0 = X, y_0 = y, X\ e\ y$ centradas o estandarizadas.
Paso 2. Para $h = 1, \ldots, a$ ($a$ es el rango de $X$).
Paso 3. $w_h = X_{h-1}^T y_{h-1} / \|X_{h-1}^T y_{h-1}\|$.
Paso 4. $t_h = X_{h-1} w_h, componente\ h\ de\ X$.
Paso 5. $p_h = X_{h-1}^T t_h / (t_h^T t_h)$.
Paso 6. $X_h = X_{h-1} - t_h p_h^T$ , Deflactamos $X$.
Paso 7. $c_h = y_{h-1}^T t_h / (t_h^T t_h)$.
Paso 8. $u_h = y_{h-1} / c_h, componente\ h\ de\ Y$.
Paso 9. $y_h = y_{h-1} - t_h c_h$ , Deflactamos $Y$.
Paso 10. *end* $h$.
Como se expresa en (2.8), la fórmula de regresión para $y$ queda

$$y \approx \hat{y} = c_1 t_1 + c_2 t_2 + c_3 t_3 + \cdots + c_a t_a \quad (2.14)$$

Pero por construcción de los $t_h$ y su relación con los $w_h$, $p_h$ y los $c_h$, podemos expresar $\hat{y}$ en terminos de las variables originales, así

$$\hat{y} = XW(P^T W)^{-1} c = XB \quad (2.15)$$

donde $B = W(P^T W)^{-1} c$, es en este caso el vector de los coeficientes de regresión PLS de $y$ sobre $X$ utilizando $h$ componentes [5,8,9].

### D. Regresión PLS, caso $L > 1$ (PLS2R)

Considere ahora el caso $L > 1$, salida múltiple, $X$ e $Y$ centradas o estandarizadas.

De las relaciones expuestas en (2.3) y (2.4), y trabajando análogamente como en PLS1R, se puede mostrar que (2.5) sujeto a $\|w\| = 1$ y $\|c\| = 1$ genera el problema de autovalores y autovectores para $w$ [5, 8]

$$X^T Y Y^T X w = \lambda w \quad (2.16)$$

Con $w$ óptimo de la forma $w = X^T u$. También, por sustituciones sucesivas se pueden obtener para $t$, $c$ y $u$, los problemas [5]

$$XX^T YY^T t = \lambda t \quad (2.17)$$
$$Y^T XX^T Y w = \lambda c \quad (2.18)$$
$$YY^T XX^T u = \lambda u \quad (2.19)$$

Aprovechando la capacidad del algoritmo NIPALS para encontrar en forma iterativa estas componentes y resolver los problemas de autovalores, generalizamos el algoritmo de PLS1R en la forma:

**Algoritmo PLS2R**

Entradas: $X$ e $Y$ centradas o estandarizadas.
Paso 1. Iniciamos con $u$ aleatorio.
Paso 2. $w = X^T u$
Paso 3. $t = Xw, t \leftarrow t/\|t\|$
Paso 4. $c = Y^T t$
Paso 5. $u = Yc, u \leftarrow u/\|u\|$
Paso 6. Repetir pasos 2 a 5 hasta convergencia.
Paso 7. Deflactar $X, Y : X \leftarrow X - tt^T X, Y \leftarrow Y - tt^T Y$

El modelo de regresión PLS descrito en (2.6) puede expresarse en la forma [5,7,8]

$$Y \approx \hat{Y} = XB \quad (2.20)$$

donde $B = W(P^T W)^{-1} C^T$ es la matriz de coeficientes de regresión. La matriz $P^T W$ es triangular superior y por tanto invertible.

La matriz $B$ se obtiene del siguiente calculo: Sabemos de (2.3) y (2.4) que la relación entre $U$ y $T$ es lineal de la forma $U = TD + H$, con $D$ una matriz diagonal y $H$ una matriz residual. Multiplicando por $W$ en ambos lados de la relación aproximada $X = TP^T$, se tiene

$$XW = TP^T W$$
$$T = XW(P^T W)^{-1} \quad (2.21)$$

Ahora volviendo a (2.4) y sustituyendo el valor de $U$ en $Y$, se tiene

$$\begin{aligned} Y &= (TD + H)Q^T + F \\ &= TDQ^T + HQ^T + F \\ &= TC^T + H^* \end{aligned} \quad (2.22)$$

donde $H^* = HQ^T + F$ es una matriz residual.

Sustituyendo (2.21) en la aproximación de $Y$ en (2.22), se tiene

$$Y \approx \hat{Y} = XW(P^TW)^{-1}C^T = XB$$

ó bien

$$Y \approx \hat{Y} = TC^T \quad (2.23)$$

### E. PLS no lineal

Siguiendo las líneas de la técnica PLS, podemos asumir dos maneras de modelar relaciones de datos no lineales a partir de la técnica PLSR [8].

La primera estrategia consiste en considerar la reformulación de la relación lineal $U = TD + H$ (entre los vectores o componentes ortogonales $t$ de $X$ y las componentes $u$ de $Y$) dada en PLS2R, por la relación no lineal

$$u = f(t) + \varepsilon = f(X, w) + \varepsilon \quad (2.24)$$

donde $f$ es una función continua que modela relaciones no lineales. Funciones polinomiales, splines y otros métodos de suavizado se han usado en la construcción de $f$.

Una segunda estrategia consiste en usar métodos Kernel en espacios de Hilbert Reproductor del Kernel (RKHS) [9]. La idea básica consiste en asumir una transformación no lineal de las variables de entrada $\{x_i\}_{i=1}^n$ en un espacio de características $F$, esto es, considerar una aplicación

$$\Phi: x_i \in \mathbb{R}^N \rightarrow \Phi(x_i) \in F \quad (2.25)$$

y luego construir un modelo PLSR en $F$ [9].

### F. Regresión Kernel PLS (KPLSR)

La técnica KPLSR se basa en mapeos del espacio original de datos $X$ a un espacio de alta dimensión $F$, luego con la aplicación del truco kernel la estimación de PLS en el nuevo espacio se reduce a cálculos de algebra lineal tan simples como en PLS lineal [8]. Como es sabido el truco kernel permite conocer el producto punto entre dos elementos $\Phi(x_i), \Phi(x_j)$ en $F$ y construir la matriz kernel $K$ sin necesidad de conocer quien es en realidad $\Phi$[9], esto es

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j), \forall x_i, x_j \in X \quad (2.26)$$

Se define así la gran matriz $K$ de productos puntos entre todos los mapeos de los puntos de datos $K = \Phi\Phi^T$ , donde $\Phi$ denota la matriz de mapeo de los elementos o datos del espacio $\{\Phi(x_i) \in F\}_{i=1}^n$.

Por lo descrito anteriormente, se deriva el algoritmo para KPLSR como una modificación del algoritmo para PLS2R en los pasos 2 y 3 [8,9].

**Algoritmo KPLSR**

Paso 1. Entradas: $K$ e $Y$ centradas.
Paso 2. Iniciamos con $u$ aleatorio.
Paso 3. $t = \Phi\Phi^T u = Ku, t \leftarrow t/\|t\|$.
Paso 4. $c = Y^T t$
Paso 5. $u = Yc, u \leftarrow u/\|u\|$.
Paso 6. Repetir pasos 2 a 5 hasta convergencia.
Paso 7. Deflactar $\Phi\Phi^T = K$

$$\Phi\Phi^T = K \leftarrow (\Phi - tt^T\Phi)(\Phi - tt^T\Phi)^T$$
$$Y \leftarrow Y - tt^TY$$

ó bien

$$K \leftarrow (I_n - tt^T)K(I_n - tt^T)$$
$$= K - tt^TK - Ktt^T + tt^TKtt^T \quad (2.27)$$

Para centrar la matriz $K$ que contiene el mapeo de datos en el espacio $F$, se aplica el procedimiento [8,9,10]

$$K = \left(I_n - \frac{1}{n}1_n1_n^T\right)K\left(I_n - \frac{1}{n}1_n1_n^T\right) \quad (2.28)$$

De manera análoga a lo expuesto en PLS2R, la matriz de coeficientes de regresión $B$ descrita en (2.20), para el caso KPLS tiene la forma [9]

$$B = \Phi^T U(T^TKU)^{-1}T^TY \quad (2.29)$$

y la predicción sobre los datos de entrenamiento está dada por

$$\hat{Y} = \Phi B$$
$$= \Phi\Phi^T U(T^TKU)^{-1}T^TY$$
$$= KU(T^TKU)^{-1}T^TY$$
$$= KG \quad (2.30)$$
$$= TT^TY$$
$$= TC^T \quad (2.31)$$

donde $G = U(T^TKU)^{-1}T^TY$ , $T = \Phi R$, $R = \Phi^T U(T^TKU)^{-1}$ [8,9].

Para hacer predicción sobre los puntos de prueba $\{x_i\}_{i=1}^n$, se utiliza la matriz de coeficientes de regresión $B$ dada en (2.29), de modo que la aproximación de las predicciones queda [9]

$$\hat{Y} = \Phi_t B = K_t U(T^TKU)^{-1}T^TY \quad (2.32)$$

De la relación (2.31) podemos interpretar los modelos KPLS como un modelo de regresión lineal, para el caso $L = 1$, en la forma [9]

$$f(\boldsymbol{x}, c) = c_1t_1(\boldsymbol{x}) + c_2t_2(\boldsymbol{x}) + \cdots + c_at_a(\boldsymbol{x})$$
$$= c^Tt(\boldsymbol{x}) \quad (2.33)$$

donde los $\{t_i(\boldsymbol{x})\}_{i=1}^a$ son la proyecciones de $\boldsymbol{x}$ en las $a$ componentes extraídas y $c$ es el vector de pesos descrito como $C$.

## III. METODOLOGÍA

### A. Elección de variables y recopilación de datos

Se seleccionaron aquellas variables consideradas de alto impacto en procesos de acreditación de alta calidad para programas de pregrado y de tipo cuantitativa. Luego se inició el proceso de recolección de datos en diferentes dependencias

de la Universidad Nacional de Colombia - Sede Manizales. Los datos recopilados corresponden a los semestres académicos 2009-I a 2012-II en los programas de Administración de Empresas Nocturna, Administración de Empresas Diurna, Administración en Sistemas Informáticos, Arquitectura, Ingeniería Civil, Ingeniería Eléctrica, Ingeniería Electrónica, Ingeniería Física, Ingeniería Industrial, Ingeniería Química y Matemáticas.

La elección sobre las variables de entrada y de salida se realizó fundamentada en los lineamientos para la acreditación de programas académicos de pregrado en Colombia establecidos por el CNA [1,2,3,4], bajo la premisa de tener en cuenta como principios de calidad **factores iniciales o propios** como: condiciones académicas, disponibilidad de recursos físicos y financieros y pertinencia social y profesional, y **factores de garantía**, **reconocimiento y aseguramiento de la calidad** como: desempeño de egresados, producción intelectual de docentes, reconocimiento de la sociedad, impacto obtenido en el medio, calidad de procesos pedagógicos, entre otros.

Así, se decide trabajar con las siguientes variables de entrada:

$x_1$ =Tasa de Absorción.
$x_2$ =Número de estudiantes matriculados provenientes de Instituciones públicas.
$x_3$ =Número de estudiantes matriculados provenientes de Instituciones privadas.
$x_4$ =Resultado en prueba de admisión de los estudiantes.
$x_5$ =Número de Estudiantes que reciben apoyo institucional (préstamos, alojamientos, alimentación).
$x_6$ =Número de Estudiantes en intercambio con universidades nacionales o extranjeras.
$x_7$ =Número de estudiantes que realizan préstamos en biblioteca.
$x_8$ =Número de estudiantes que acceden a consulta en las bases de datos SINAB.
$x_9$ =Número de Estudiantes vinculados como monitores.
$x_{10}$ =Número de Estudiantes que reservan cupo.
$x_{11}$ =Número de Estudiantes que cancelan semestre.
$x_{12}$ =Número de docentes vinculados al programa, cuyo título superior obtenido es el de maestría.
$x_{13}$ =Número de docentes con título de doctorado vinculados al programa.
$x_{14}$ =Número de docentes con dedicación medio tiempo y tiempo completo.
$x_{15}$ =Número de docentes con dedicación exclusiva.
$x_{16}$ =Número de convenios del programa establecidos con el sector productivo para efectos de prácticas.
$x_{17}$ =Número de semilleros de investigación en el programa.
$x_{18}$ =Número de estudiantes vinculados a semilleros de investigación.
$x_{19}$ =Número de docentes que lideran semilleros de investigación.
$x_{20}$ =Número de grupos de investigación reconocidos.
$x_{21}$ =Número de docentes que lideran grupos de investigación en el programa.
$x_{22}$ =Productividad académica de docentes.

El número de variables de entrada varía de acuerdo al programa analizado. En los programas Administración de Empresas Nocturna y Diurna no se analiza la variable de entrada $x_{14}$, y se une $x_{20}$ con $x_{21}$. En el programa de Administración en Sistemas Informáticos se unen $x_{20}$ y $x_{21}$. En el programa de Arquitectura e Ingeniería Industrial no se analiza $x_{16}$, y se une $x_{20}$ con $x_{21}$. En el programa de Ingeniería Civil no se analiza $x_{16}$, y se unen $x_{19}$, $x_{20}$ y $x_{21}$. En los programas de Ingeniería Eléctrica e Ingeniería Electrónica no se analizan $x_6$ y $x_{16}$ , y se unen $x_{20}$ y $x_{21}$. En el programa de Ingeniería Física no se analizan $x_6$, $x_9$ y $x_{16}$, y se unen $x_{20}$ y $x_{21}$. En el programa de Ingeniería Química no se analizan $x_6$, $x_{14}$ y $x_{16}$. Finalmente en el programa de Matemáticas no se analizan $x_6$ y $x_{16}$, se unen $x_{17}$ con $x_{19}$ y $x_{20}$ con $x_{21}$. Las razones por las cuales no se analizan o se excluyen ciertas variables del estudio se deben a que estas no aplican para el programa o no se tiene información completa de ellas en los periodos analizados; y la unión de variables se da porque estas comparten la misma información, en cuyo caso se decide dejar una en representación de todas.

Las variables de salida analizadas en todos los programas son:

$y_1$ = Resultados de estudiantes en pruebas Saber – Pro.
$y_2$ = Promedio académico estudiantil.
$y_3$ = Número de estudiantes que desertan del programa.

Para la recolección de datos se usaron las siguientes fuentes: http://portal.manizales.unal.edu.co/planeacion/index.php/estadisticas, Plataforma SARA, Sistema de Información de Evaluación Educativa (ICFES), e Información suministrada por las siguientes dependencias de la Universidad Nacional de Colombia – Sede Manizales: Departamento de Planeación, Dirección de investigación sede Manizales (DIMA), Facultades a las cuales pertenecen los programas estudiados, ORI, CEUNAL, Bienestar Universitario, Oficina de registro y matrícula y Dirección Académica.

### B. Adecuación de la base de datos

Se realizan ajustes sobre los datos y las observaciones, según el caso, entre los que se encuentran: ajustes de escala, cambio de representación numérica, duplicación de datos y eliminación de datos faltantes. De modo que para todos los programas se consideraron 8 observaciones, correspondientes a los semestres 2009-I a 2012-II.

### C. Construcción y estructuración de los modelos

Para el caso $L = 1$, en cada programa se presentan modelos iniciales de regresión lineal múltiple usando las técnicas mínimos cuadrados (LS) y mínimos cuadrados parciales caso univariado (PLS1R), con las estructuras descritas en el capítulo anterior.

Para el caso $L > 1$, en cada programa se presentan modelos de regresión multivariado usando las técnicas mínimos cuadrados parciales (PLS) para el caso lineal y Kernel PLS

para el caso no lineal, con las estructuras descritas en el capítulo anterior.

Los parámetros para cada modelo se obtienen a partir de algoritmos diseñados en MATLAB, teniendo en cuenta las versiones NIPALS y las teorías que cada técnica implementa.

### D. Implementación, ajuste y validación de los modelos

Se implementan los modelos descritos anteriormente en cada programa académico sobre la base de datos recolectada, y se realizan procesos de validación usando el método de validación cruzada dejando uno por fuera (*en inglés Cross Validation-Leave one out (LOO)*); se seleccionan los de mejor comportamiento predictivo.

Para el caso de salida múltiple se ajustan aquellos modelos cuyo comportamiento predictivo no es el deseado, a partir de técnicas no lineales como Kernel PLS, creando modelos de regresión no lineales (KPLSR) que contribuyen en la mejora del aspecto predictivo comparados con los obtenidos en PLSR. La función kernel usada aquí es la Kernel Gaussiana

$$K(x_i, x_j) = e^{-\left(\frac{||x_i - x_j||^2}{d}\right)} \tag{3.1}$$

## IV. RESULTADOS Y DISCUSIÓN

Se presentan los resultados obtenidos en la implementación, ajuste y validación de los modelos LS, PLS1R, PLS2R y KPLSR con su respectiva discusión, tomando como ejemplo el programa Administración de Empresas Nocturno. Para el resto de programas se hace un trabajo análogo.

Los modelos con $L = 1$ tienen en cuenta como variable de salida $y_1$ y los de salida múltiple las tres variables $y_1$, $y_2$ e $y_3$, descritas en la metodología.

### A. Modelo LS

Aplicando la técnica de mínimos cuadrados se obtiene el modelo

$$\hat{Y} = \begin{bmatrix} -6{,}28E-02 \\ 1{,}63E-01 \\ -3{,}85E-02 \\ -3{,}83E-02 \\ -7{,}06E-02 \\ 9{,}51E-02 \\ -5{,}12E-02 \\ 9{,}02E-02 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ x_7 \\ x_8 \\ x_{13} \\ x_{16} \\ x_{17} \\ x_{20} \end{bmatrix} \tag{4.1}$$

Se realizan las validaciones pertinentes, tomando siete datos de entrenamiento y uno de prueba. Se implementa el modelo sobre estos siete datos y se predice el excluido, obteniendo como resultado

| OBS. | Real | Predicción | Errores |
|---|---|---|---|
| 1 | 105,400 | 103,524 | 1,78E-02 |
| 2 | 105,400 | 108,171 | 2,63E-02 |
| 3 | 105,280 | 105,669 | 3,70E-03 |
| 4 | 105,280 | 104,553 | 6,90E-03 |
| 5 | 103,100 | 107,468 | 4,24E-02 |
| 6 | 105,700 | 100,442 | 4,97E-02 |
| 7 | 106,200 | 104,489 | 1,61E-02 |
| 8 | 105,000 | 107,770 | 2,64E-02 |

Tabla 4.1. Predicciones del modelo LS y errores de predicción para $\boldsymbol{y_1}$.

### B. Modelo PLS1R

Aplicando la técnica PLS1, se obtiene el modelo de regresión

$$\hat{Y} = 118{,}7403 + \begin{bmatrix} -5{,}204E-04 \\ 2{,}116E-02 \\ 2{,}914E-02 \\ -1{,}087E-02 \\ -1{,}238E-02 \\ -1{,}077E-02 \\ -2{,}846E-02 \\ -3{,}020E-02 \\ -2{,}815E-02 \\ -3{,}731E-02 \\ 5{,}826E-02 \\ -5{,}426E-03 \\ -7{,}738E-03 \\ -6{,}187E-04 \\ -2{,}608E-02 \\ 3{,}115E-03 \\ -4{,}258E-02 \\ -2{,}094E-03 \\ -9{,}450E-04 \\ -1{,}665E-02 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{20} \end{bmatrix} \tag{4.2}$$

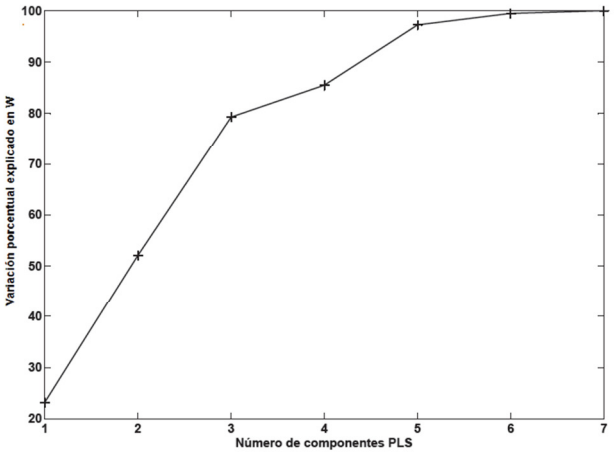


Fig. 4.1. Influencia de las variables latentes o componentes PLS sobre $Y$.

| OBS. | Real | Predicción | Errores |
|---|---|---|---|
| 1 | 105,400 | 103,927 | 1,40E-02 |
| 2 | 105,400 | 106,320 | 8,73E-03 |
| 3 | 105,280 | 105,240 | 3,78E-04 |
| 4 | 105,280 | 105,838 | 5,30E-03 |
| 5 | 103,100 | 105,517 | 2,34E-02 |
| 6 | 105,700 | 102,899 | 2,65E-02 |
| 7 | 106,200 | 105,241 | 9,03E-03 |
| 8 | 105,000 | 105,702 | 6,69E-03 |

Tabla 4.2. Predicciones del modelo PLS1R y errores de predicción para $\mathbf{y_1}$.

Para el programa de Admistración de Empresas Nocturna el modelo LS, presenta en el proceso de validación un error máximo del 4,97 %, lo que implica un porcentaje mínimo de predicción del 95,03% (de aquí en adelante a esto se le llamará porcentaje de predicción). El modelo PLS1R mejora los resultados predictivos con un error máximo de 2,64%, esto es, con porcentaje de predicción del 97,36%. Significa que si se espera un valor real de 102 puntos el valor predicho con la técnica LS estaría entre 96,9 y 106,9 puntos (aproximadamente), mientras que con la técnica PLS1 oscila entre 99,3 y 104,04 puntos, generando una ganancia en precisión. Sin embargo, ambos modelos son considerados adecuados para efectos predictivos.

Ahora bien, si nos concentramos en el aspecto explicativo, las diferencias son notables con PLS1R. La técnica LS usa como criterio de optimización el error mínimo entre $Y$ y $Y$ aproximado, dejando a un lado el análisis de la influencia de las variables de entrada sobre la variable de salida $Y$. Por ejemplo, para el caso que nos ocupa, la técnica LS permite obtener un modelo que se expresa en 8 variables originales de las 20 utilizadas inicialmente, pero no hay garantía que sean estas 8 variables las que mejor explican la variabilidad de $Y$, más aún no se garantiza que estas estén o no correlacionadas. Por su parte, la técnica PLS1R además del aspecto predictivo, resuelve en gran parte el aspecto explicativo y de paso el de multicolinealidad, tal como se expone en (2.8). De esta manera para el caso que nos ocupa los $t_i's$ toman la forma

$$t_i = p_{1i}x_1 + p_{3i}x_2 + \cdots + p_{20i}x_{20}, \;\; i = 1, \ldots, 7 \qquad (4.3)$$

donde $p_{1i}, \ldots, p_{20i}$ son los elementos de la primera columna de $P^T$ de la ecuación (2.3). El porcentaje explicativo de cada $t_i$ en $Y$ se muestra en la Figura 4.1.

Con la relación (4.3) se exploran las variables originales más correlacionadas con las componentes ortogonales de mayor porcentaje explicativo en $Y$ $(cor(x_i, t_j))$, y de esta transitividad sabremos qué variables originales aportan más en la variabilidad de $Y$ [5]. Las variables originales que cumplan con estas condiciones de ahora en adelante se dirán las más influyentes en $Y$.

En este sentido, se indagó sobre las variables originales que más influyeron en el comportamiento de los resultados Saber-Pro, durante los semestres 2009-I a 2012-II, en cada uno de los 11 programas académicos, encontrando con mayor frecuencia las variables:

- Número de grupos de investigación y número de docentes que lideran grupos de investigación.
- Productividad Académica.
- Número de estudiantes que realizan préstamos en biblioteca.
- Número de estudiantes que acceden a consultas en las bases de datos Sinab.
- Número de estudiantes vinculados a semilleros de investigación.

Cabe aclarar que el tipo de relación explicativa que aquí se menciona no es necesariamente directa; sin embargo, es información importante para tener en cuenta en la toma de decisiones. De otra parte, note que la mayoría de las variables explicativas citadas corresponden a procesos que depende del aporte y trabajo de los estudiantes, el trabajo docente y a procesos académicos e investigativos liderados por la institución; lo que refleja el grado de contribución de la Universidad Nacional de Colombia - Sede Manizales en aporte de valor agregado para los estudiantes de estos programas.

### C. Modelo PLS2R

Se recuerda que aquí las variables de salida corresponden a resultados en pruebas Saber-Pro, promedio académico y deserción. Aplicando la técnica PLS se obtuvo el modelo

$$\hat{Y} = \begin{bmatrix} 118{,}7403213 \\ 3{,}026418071 \\ -157{,}3675828 \end{bmatrix} + B^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{20} \end{bmatrix} \qquad (4.4)$$

Donde $B$ esta dada por la matriz:

$$B = \begin{bmatrix} -5{,}204E-04 & -2{,}398E-05 & 6{,}969E-03 \\ 2{,}116E-02 & 8{,}862E-04 & -4{,}913E-01 \\ 2{,}914E-02 & 1{,}161E-03 & -2{,}675E-01 \\ -1{,}087E-02 & 6{,}716E-04 & 2{,}750E-01 \\ -1{,}238E-02 & -1{,}565E-03 & 8{,}860E-01 \\ -1{,}077E-02 & 2{,}994E-05 & -7{,}642E-02 \\ -2{,}846E-02 & 4{,}305E-04 & 3{,}173E-02 \\ -3{,}020E-02 & -7{,}472E-04 & 6{,}597E-01 \\ -2{,}815E-02 & -5{,}922E-04 & 3{,}846E-01 \\ -3{,}731E-02 & -1{,}484E-04 & 4{,}855E-02 \\ 5{,}826E-02 & 8{,}864E-04 & -6{,}212E-01 \\ -5{,}426E-03 & -5{,}124E-05 & 1{,}208E-02 \\ -7{,}738E-03 & -3{,}307E-04 & 1{,}963E-01 \\ -6{,}187E-04 & 1{,}681E-05 & -1{,}909E-03 \\ -2{,}608E-02 & -6{,}095E-04 & 3{,}162E-01 \\ 3{,}115E-03 & -4{,}163E-04 & -1{,}412E-01 \\ -4{,}258E-02 & -1{,}296E-03 & 6{,}138E-01 \\ -2{,}094E-03 & 2{,}809E-04 & -7{,}363E-02 \\ -9{,}450E-04 & 1{,}180E-04 & -3{,}080E-02 \\ -1{,}665E-02 & -2{,}899E-04 & 2{,}459E-01 \end{bmatrix} \quad (4.5)$$

| PREDICCION | OBSERVACIONES | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| YERRORES | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $y_1$ | 105,400 | 105,400 | 105,280 | 105,280 | 103,100 | 105,700 | 106,200 | 105,000 |
| $y_1 p$ | 103,927 | 106,320 | 105,240 | 105,838 | 105,517 | 102,899 | 105,241 | 105,702 |
| $e_1$ | 3,78E-04 | 8,73E-03 | 3,78E-04 | 5,30E-03 | 2,34E-02 | 2,65E-02 | 9,03E-03 | 6,69E-03 |
| $y_2$ | 3,699 | 3,663 | 3,589 | 3,565 | 3,596 | 3,572 | 3,605 | 3,597 |
| $y_2 p$ | 3,674 | 3,667 | 3,563 | 3,604 | 3,654 | 3,528 | 3,605 | 3,573 |
| $e_2$ | 7,36E-03 | 1,12E-03 | 7,36E-03 | 1,10E-02 | 1,60E-02 | 1,23E-02 | 8,69E-05 | 6,68E-03 |
| $y_3$ | 17,000 | 24,000 | 22,000 | 45,000 | 42,000 | 26,000 | 24,000 | 36,000 |
| $y_3 p$ | 31,606 | 18,390 | 42,378 | 20,338 | 7,214 | 59,917 | 24,875 | 48,500 |
| $e_3$ | 9,26E-01 | 2,34E-01 | 9,26E-01 | 5,48E-01 | 8,28E-01 | 1,30 | 3,65E-02 | 3,47E-01 |

Tabla 4.3. Predicciones ($y_i p$) del modelo PLS2R y errores de predicción ($e_i$).

### 4.3.1. Exploración sobre el comportamiento predictivo de las variables de salida $y_2$ e $y_3$ en forma independiente, usando PLS1R.

| OBS. | Real | Predicción | Errores |
|---|---|---|---|
| 1 | 3,699 | 3,6742 | 6,72E-03 |
| 2 | 3,663 | 3,6671 | 1,12E-03 |
| 3 | 3,589 | 3,5626 | 7,36E-03 |
| 4 | 3,565 | 3,6041 | 1,10E-02 |
| 5 | 3,596 | 3,6535 | 1,60E-02 |
| 6 | 3,572 | 3,5279 | 1,23E-02 |
| 7 | 3,605 | 3,6047 | 8,69E-05 |
| 8 | 3,597 | 3,5730 | 6,68E-03 |

Tabla 4.4. Predicciones del modelo PLS1R y errores de predicción para $y_2$ (promedio académico).

| OBS. | Real | Predicción | Errores |
|---|---|---|---|
| 1 | 17 | 31,606 | 8,59E-01 |
| 2 | 24 | 18,390 | 2,34E-01 |
| 3 | 22 | 42,377 | 9,26E-01 |
| 4 | 45 | 20,338 | 5,48E-01 |
| 5 | 42 | 7,213 | 8,28E-01 |
| 6 | 26 | 59,91 | 1,30 |
| 7 | 24 | 24,875 | 3,65E-02 |
| 8 | 36 | 48,499 | 3,47E-01 |

Tabla 4.5. Predicciones del modelo PLS1R y errores de predicción para $y_3$ (deserción)

Como puede apreciarse en la Tabla 4.3, el modelo PLS2R muestra buen comportamiento predictivo en las dos primeras variables de salida Saber-Pro (97,36%) y Promedio Académico (99,2%), pero no para la variable Deserción. Este comportamiento quizás se deba a dos cosas: la primera, a su fuerte variabilidad con el paso del tiempo, lo que implica relacionarla con un patrón no lineal; la segunda, por acomodamiento de pesos, es decir, que los pesos generados en $B$ para conseguir buen comportamiento predictivo en las variables Saber-Pro y Promedio Académico, afectaron el comportamiento predictivo para la variable Deserción, sugiriendo un tratamiento individual. En realidad, una exploración realizada con cada variable de salida aplicando PLS1R descarta la segunda posibilidad (compare las Tablas 4.2, 4.4 y 4.5 con la Tabla 4.3).

De la relación (2.23) se procede análogamente como en PLS1R y se determinan las variables de entrada que mejor explican el comportamiento de los resultados Saber-Pro, Promedio Académico y Deserción en cada programa durante los semestres 2009-I a 2012-II, encontrándose con mayor frecuencia las variables:

- Número de estudiantes que realizan préstamos en biblioteca.
- Número de grupos de investigación y número de docentes que lideran grupos de investigación.
- Número de docentes que laboran con dedicación exclusiva.
- Número de estudiantes vinculados a semilleros de investigación.
- Número de docentes con título de doctorado vinculados al programa.
- Número de estudiantes que realizan consultas en bases de datos SINAB.

Esta información es aún más interesante para efectos de toma de decisiones y aporte de valor agregado, por cuanto se mira la influencia sobre tres variables de salida, consideradas de gran impacto en calidad educativa.

### D. Modelos KPLSR

Se aplica la técnica KPLSR (caso $L = 1$) para el programa Administración de Empresas Nocturna, tomando como variable de salida Deserción, la cual presentó dificultad en predicción usando los modelos lineales.

El modelo toma la estructura dada en (2.30) o bien $Y(x, g) = \sum_{i=1}^{n} g_i K(x, x_i)$, donde $g$ es el vector de pesos

$$g = \begin{bmatrix} -1{,}1495 \\ -1{,}2603 \\ 22{,}1600 \\ 19{,}0234 \\ 3{,}0173 \\ 0{,}8141 \\ 13{,}1899 \end{bmatrix} \quad (4.6)$$

| OBS. | Real | Predicción | Errores |
|---|---|---|---|
| 1 | 17 | 14,000 | 1,76E-01 |
| 2 | 24 | 15,875 | 3,39E-01 |
| 3 | 22 | 36,938 | 6,79E-01 |
| 4 | 45 | 55,500 | 2,33E-01 |
| 5 | 42 | 39,250 | 6,55E-02 |
| 6 | 26 | 29,000 | 1,15E-01 |
| 7 | 24 | 34,000 | 4,17E-01 |
| 8 | 36 | 36,094 | 2,60E-03 |

Tabla 4.6. Predicciones del modelo KPLSR ($\boldsymbol{L=1}$) y errores de predicción en $\boldsymbol{y_3}$.

| OBS. | Errores PLS1R | Errores KPLSR |
|---|---|---|
| 1 | 8,59E-01 | 1,76E-01 |
| 2 | 2,34E-01 | 3,39E-01 |
| 3 | 9,26E-01 | 6,79E-01 |
| 4 | 5,48E-01 | 2,33E-01 |
| 5 | 8,28E-01 | 6,55E-02 |
| 6 | 1,30 | 1,15E-01 |
| 7 | 3,65E-02 | 4,17E-01 |
| 8 | 3,47E-01 | 2,60E-03 |

Tabla 4.7. Comparación de errores PLS1R vs KPLSR ($\boldsymbol{L=1}$).

Ahora se aplica la técnica KPLSR para el caso $L>1$, tomando como variables de salida: Resultados Saber-Pro, Promedio Académico y Deserción. El modelo para este caso toma la forma $\hat{Y}=KG$, donde $K$ es la matriz Kernel y $G$ es la matriz de coeficientes

$$G=\begin{bmatrix} -81{,}927 & -3{,}568E-01 & 523{,}765 \\ -81{,}927 & -3{,}932E-01 & 530{,}839 \\ -82{,}046 & -4{,}655E-01 & 528{,}429 \\ -82{,}046 & -4{,}899E-01 & 551{,}849 \\ -84{,}225 & -4{,}593E-01 & 548{,}713 \\ -81{,}624 & -4{,}833E-01 & 532{,}707 \\ -81{,}117 & -4{,}499E-01 & 530{,}503 \\ -82{,}355 & -4{,}581E-01 & 542{,}879 \end{bmatrix} \quad (4.7)$$

| PREDICCION Y ERRORES | OBSERVACIONES | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $y_1$ | 105,400 | 105,400 | 105,280 | 105,280 | 103,100 | 105,700 | 106,200 | 105,000 |
| $y_1p$ | 105,650 | 105,494 | 105,483 | 103,186 | 103,475 | 105,200 | 105,200 | 105,020 |
| $e_1$ | 2,37E-03 | 8,89E-04 | 1,93E-03 | 1,99E-02 | 3,64E-03 | 4,73E-03 | 9,42E-03 | 1,86E-04 |
| $y_2$ | 3,699 | 3,663 | 3,589 | 3,565 | 3,596 | 3,572 | 3,605 | 3,597 |
| $y_2p$ | 3,764 | 3,650 | 3,540 | 3,543 | 3,555 | 3,552 | 3,582 | 3,597 |
| $e_2$ | 1,75E-02 | 3,50E-03 | 1,36E-02 | 6,16E-03 | 1,15E-02 | 5,47E-03 | 6,50E-03 | 7,95E-05 |
| $y_3$ | 17,000 | 24,000 | 22,000 | 45,000 | 42,000 | 26,000 | 24,000 | 36,000 |
| $y_3p$ | 9,750 | 15,875 | 36,250 | 59,875 | 39,000 | 34,000 | 39,000 | 35,934 |
| $e_3$ | 4,26E-01 | 3,39E-01 | 6,48E-01 | 3,31E-01 | 7,14E-02 | 3,08E-01 | 6,25E-01 | 1,84E-03 |

Tabla 4.8. Predicciones del modelo KPLSR ($\boldsymbol{L>1}$) y errores de predicción en $\boldsymbol{Y}$.

Este proceso de implementación y validación se realiza en forma análoga para el resto de programas en estudio.

Para el caso $L=1$ con Deserción como variable de salida, los resultados en predicción mejoran considerablemente con respecto a lo obtenido con PLS1R (ver Tabla 4.7). Note que el error máximo con PLS1R es de 1,30 aproximadamente, mientras que con KPLSR es de 0,115, una disminución muy interesante. Exploraciones realizadas con esta técnica permiten conjeturar que estos resultados predictivos podrían mejorar si contáramos con más observaciones en el estudio. Fíjese que esta técnica aplicada con 8 observaciones genera modelos cuyos residuales son muy pequeños del orden de $10^{-15}$, pero como en el proceso de validación cruzada (LOO) debemos prescindir de una de las observaciones y trabajar solo con 7; obtenemos resultados con residuales no tan pequeños (ver Figura 4.2). Por tanto las predicciones con estos parámetros obtenidos no van a ser las deseadas. Sin embargo, las mejoras en el aspecto predictivo comparadas con los modelos PLS1R son notorias.

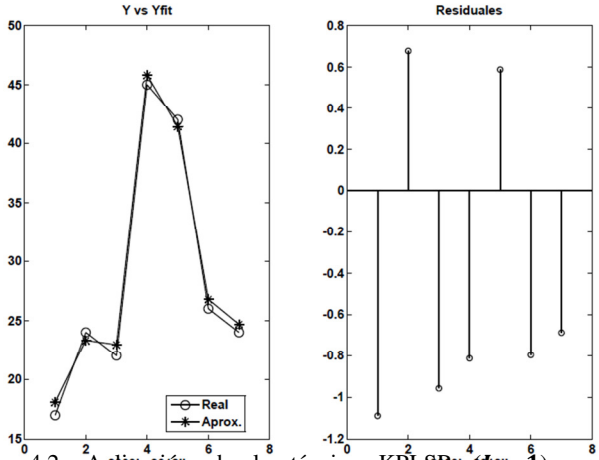


Fig. 4.2. Aplicación de la técnica KPLSR ($\boldsymbol{L=1}$), con 7 observaciones.

En lo que respecta a la aplicación de la técnica KPLSR (salida múltiple ó $L>1$), los resultados predictivos comparados con PLS2R son interesantes (compare las Tablas 4.3 y 4.8). Note que las predicciones en las variables de salida con más dificultad predictiva mejoran, característica que se mantiene en todos los programas académicos estudiados. Por ejemplo, el error máximo en la predicción de la variable deserción para el programa Administración de empresas nocturna con la técnica PLS2R es de 1,30, y con la técnica KPLSR se reduce a 0,426 aproximadamente, es decir, en más del 50 %.

Observe que cuando se usó PLS2R con la intención de mejorar los errores de predicción para la variable Deserción, recurriendo a un tratamiento individual con PLS1R, los resultados se mantuvieron; ahora, aplicando la técnica KPLSR con $L=1$ ó $L>1$ los resultados varían notoriamente (comparar Tablas 4.6 y 4.8), es decir, que con esta técnica es más provechoso para objetos predictivos un tratamiento individual. Las mejoras en el resto de programas académicos también son notorias.

En cuanto a la interpretación de los modelos KPLSR, se recurre a la aproximación $Y=TC^T$. Para el caso $L=1$, $Y$ queda expresada en la forma de la ecuación (2.33); para el caso $L>1$, cada componente de $Y$ se puede ver como una combinación lineal de las columnas de $T$, donde cada $t_i$ es una función no lineal que depende de $x$.

A diferencia de la técnica PLS2R, sólo se puede explorar la relación de $Y$ con las componentes latentes, en el sentido de conocer cuál es el porcentaje explicativo de cada una de ellas

en $Y$, pero se dificulta conjeturar sobre la influencia de las variables originales en $Y$, ya que ahora los $t_i's$ son no lineales.

## V. CONCLUSIONES

- En este trabajo se encontró que los modelos lineales PLS1R presentan mejor comportamiento tanto en lo predictivo como en lo explicativo, comparado con los modelos de regresión LS.

- Los modelos lineales PLSR presentan dificultades en predicción cuando el comportamiento o variabilidad de una de las variables de salida es muy fuerte con el paso del tiempo.

- Los modelos lineales PLSR presentan su fortaleza en el aspecto explicativo; por tanto, pueden tomarse como modelos de valor agregado, para estudios de calidad educativa.

- Los modelos no lineales KPLSR presentaron mejoras en predicción con respecto a los modelos lineales PLSR, que podrían ser mejor si se consideran mayor número de observaciones.

- En el ámbito explicativo, las variables asociadas a los factores estudiantes, docencia, procesos académicos e investigación que con mayor frecuencia influyeron en el comportamiento de las variables resultados Saber-Pro, promedio académico y deserción, durante los periodos 2009-I a 2012-II, fueron las relacionadas con: estudiantes que realizan préstamos en biblioteca y consultas en bases de datos SINAB, número de grupos de investigación y docentes que lideran grupos de investigación, estudiantes vinculados a semilleros de investigación, docentes que laboran con dedicación exclusiva y docentes con título de doctorado.

## REFERENCIAS

[1] CNA, Indicadores para la autoevaluación con fines de acreditación institucional, Bogotá: Serie documentos CNA, 2006, pp. 12-52.

[2] CNA, Guía para la evaluación externa con fines de acreditación institucional, Bogotá: Serie documentos CNA, 2006, pp. 8-12.

[3] CNA, Lineamientos para la acreditación de programas de pregrado, Bogotá: Serie documentos CNA, 2013, pp. 6-20.

[4] CNA, Guía de autoevaluación con fines de acreditación de programas de pregrado, Bogotá: Serie documentos CNA, 2013, pp. 1-8.

[5] M. Tenenhaus, La régression PLS théorie et pratique, Paris: Editions Technip, 1998, pp. 61-192.

[6] G. Mateos, A. Morales, "Partial Least Square (PLS) Methods: Origins, Evolution and Application to Social Sciences", pp. 1-13, 2011.

[7] A. Höskuldsson, "PLS Regression Methods", Journal of Chemometrics, vol. 2, pp. 211-228, 1998.

[8] R. Rosipal, N. Krämer, "Overview and Recent Advances in Partial Least Squares", Springer, pp. 34-49, 2006.

[9] R. Rosipal, L. J. Trejo, "Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space", Journal of Machine Learning Research 2, pp. 97-123, 2001.

[10] M. Momma, K. P. Bennett, "Sparse Kernel Partial Least Squares Regression", Springer, pp. 21-49, 2003.

[11] G. Blanchard, N. Krämer, "Kernel Partial Least Squares is Universally Consistent", vol. 9 of JMLR: W&CP 9, pp. 57-64, 2010.

[12] K. P. Bennett and M. J. Embrechts, "An Optimization Perspective on Kernel Partial Least Squares Regression", vol. 190, IOS Press Amsterdam, pp. 8-11, 2003.