Regresión por Mínimos Cuadrados Parciales

Autor: Lic. Dindo Valdéz Blanco

1. Introducción

La regresión por mínimos cuadrados parciales, denominado regresión PLS (partial least squares), es una técnica que combina dos técnicas del análisis multivariante; el análisis de componentes principales y la regresión lineal múltiple.

La regresión PLS se utiliza generalmente en dos situaciones: cuando se tiene un gran número de variables predictoras, el número de variables independientes puede ser incluso mayor al número de observaciones, y/o cuando existe multicolinealidad entre las variables predictoras.

2. El Análisis Multivariante

Se considera el caso de un modelo lineal con una variable dependiente y m variables independientes, representados por la ecuación.

$$Y_{n\times 1} = X_{m\times n}\cdot \beta_{m\times 1} + E_{n\times 1} \qquad (1)$$

Donde Y representa la variable dependiente, X representa la matriz de variables independientes, β es el vector de coeficientes y E el vector de errores o residuos.

Respecto ala ecuación (1) se pueden considerar dos situaciones: Si $n > m$, entonces por lo general existe una única solución de mínimos cuadrados para la ecuación (1). Si $n < m$. Entonces no existe solución para la ecuación (1), en vista que la matriz X puede ser singular. O en su defecto existen infinitas soluciones de mínimos cuadrados.

3. El método de Mínimos Cuadrados Parciales (PLS)

La suposición básica de la regresión PLS es que el sistema depende de un número pequeño de variables instrumentales llamadas variables latentes. Este concepto es similar al de componentes principales. Las variables latentes son estimadas como combinaciones lineales de las variables observadas, como se explica más adelante. En los modelos PLS, se establece una representación de la matriz **X** en término de dichas variables latentes:

$$X_{n\times m} = \mathrm{T}_{\mathrm{n\times a}}\mathrm{P}^{\mathrm{T}}_{\mathrm{a\times n}} + E_{n\times m} \qquad (2)$$

donde **T** representa los "scores" (término que puede ser traducido como "resultados"); y la matriz **P** es denominada "loadings" (término que puede ser traducido como "cargas"). De esta manera la matriz **X** queda descompuesta en un número de "variables latentes", cada una caracterizada por un vector **t** y un vector $\mathbf{p}^{\mathrm{T}}$.

De esta forma, es posible representar la matriz **X** por una matriz **T** con un número menor de columnas. Esta descomposición se muestra en la ecuación (3).

$$\begin{aligned} X_{n\times m} &= \mathrm{T}_{\mathrm{n\times a}}\mathrm{P}^{\mathrm{T}}_{\mathrm{a\times n}} + E_{n\times m} \\ &= \mathrm{t}(1)_{\mathrm{n\times 1}}\cdot \mathrm{p}(1)^{\mathrm{T}}_{\mathrm{1\times n}} + \mathrm{t}(2)_{\mathrm{n\times 1}}\cdot \mathrm{p}(2)^{\mathrm{T}}_{\mathrm{1\times n}} + \cdots \\ &\cdots + \mathrm{t}(\mathrm{a})_{\mathrm{n\times 1}}\cdot \mathrm{p}(\mathrm{a})^{\mathrm{T}}_{\mathrm{1\times n}} + E_{n\times m} \quad ; \quad \mathrm{a} < \mathrm{m} \end{aligned} \qquad (3)$$

Si se incluyen todas las variables latentes, el error es cero (**E** = **0**).

El modelo PLS se desarrolla de modo que las primeras variables latentes ($\mathbf{t}_{(1)}$, $\mathbf{t}_{(2)}$,...) sean las más importantes para explicar el vector **Y** en la muestra. El número de variables latentes necesarias para explicar la matriz **X** es una medida de la complejidad del modelo. Otros vectores calculados durante la etapa de construcción del modelo son el vector **w** (llamado "pesos" de **X**), y el vector **b** (denominado "sensibilidades"), La relación entre el vector **Y** y la matriz **T** es:

$$Y_{n\times 1} = \mathrm{T}_{n\times a}\,\mathrm{b}_{a\times 1} + F_{n\times m} \qquad (4)$$

Donde **b** se calcula para minimizar los errores ***F***. El vector **Y** es estimado usando los coeficientes de **b** previamente estimados por mínimos cuadrados:

$$\hat{Y}_{n\times 1} = \mathrm{T}_{n\times a}\,\mathrm{b}_{a\times 1} \qquad (5)$$

Si se toman en cuenta todas las variables latentes (a=m), los coeficientes del vector **b** son idénticos a los coeficientes del modelo de regresión lineal múltiple:

$$\mathrm{b}_{a\times 1} = \hat{\beta} = (X^T X)^{-1} X^T Y \qquad (6)$$

4. Ventajas y desventajas del método PLS

Como se ha indicado, el método PLS obtiene a partir de la matriz **X**, una matriz **T** cuyos vectores son linealmente independientes, definiendo un sistema ortogonal. De tal forma que, en los casos en que existan un número mayor de variables independientes en relación al número de observaciones (m>n) se produce una reducción del modelo. Por otro lado, en los casos en que exista colinealidad o redundancia entre las variables, la matriz **T** se usa para reducir dichas variables o sintetizarlas. Por consecuencia es posible minimizar el riesgo de cometer un error éstadístico al descartar información importante.

Una desventaja es que la regresión PLS es un modelo correlativo y no causal, en el sentido de que los modelos obtenidos no ofrecen información fundamental acerca del fenómeno estudiado, puesto que no se trabaja con las variables originales.

5. Regresión por Componentes Principales

La regresión por componentes principales consiste de dos etapas. Primero se realiza el análisis de componentes principales de la matriz de datos **X**, y luego se utilizan estos componentes principales como las variables independientes de la función de regresión final que se construye utilizando la técnica de mínimos cuadrados entre los datos proyectados y la variable respuesta **Y.** El hecho que los componentes principales son ortogonales resuelve el problema de multicolinealidad.

La desventaja de este método radica en que los componentes principales son calculados para explicar a **X** y no toman en cuenta a la variable dependiente, puesto que estas se calculan solo con la matriz de datos de **X**. Por lo tanto nada garantiza que los componentes principales los cuales "explican" **X**, también sean relevantes para explicar a **Y**.

6. Ejemplo de aplicación

Para ilustrar las diferencias entre la regresión PLS y la regresión por componentes principales utilizaremos datos simulados con 9 observaciones, 1 variable dependiente y 9 variables independientes. La matriz de correlaciones entre las variables se muestra a continuación:

Cuadro 1. Matriz de Correlaciones

	y	x1	x2	x3	x4	x5	X6	x7	x8	x9
y	1.00	0.27	-0.55	0.67*	-0.20	-0.08	-0.17	0.07	0.32	0.27
x1		1.00	-0.39	0.59	-0.67	-0.22	-0.03	0.42	0.39	1.00**
x2			1.00	-0.42	0.27	-0.12	0.14	0.11	-0.03	-0.39
x3				1.00	-0.21	0.24	0.10	0.47	0.52	0.59
x4					1.00	0.79*	0.61	0.30	0.14	-0.67*
x5						1.00	0.75*	0.54	0.41	-0.22
x6							1.00	0.67*	0.63	-0.03
x7								1.00	0.88**	0.42
x8									1.00	0.39
x9										1.00

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

Se observa la presencia de multicolinealidad entre las variables explicativas y al mismo tiempo se tienen pocas observaciones. Aplicando la técnica de componentes principales para reducir la matriz X, se tienen los siguientes resultados en el Cuadro 2.

Cuadro 2. Resultados del Método de Componentes Principales

Componente	Autovalores iniciales	% de la varianza	% acumulado	Autovalores seleccionados	% de la varianza	% acumulado
1	3.617	40.19	40.2	3.617	40.2	40.2
2	3.305	36.72	76.9	3.305	36.7	76.9
3	1.095	12.17	89.1	1.095	12.2	89.1
4	0.468	5.19	94.3			
5	0.280	3.11	97.4			
6	0.185	2.06	99.4			
7	0.051	0.56	100.0			
8	0.000	0.00	100.0			
9	0.000	0.00	100.0			

Por tal razón se eligen los primeros 3 componentes y se aplica la regresión lineal múltiple para explicar a la variable respuesta en función de estas tres variables sintéticas, el cuadro siguiente muestra los resultados de la regresión múltiple con los tres componentes:

Cuadro 3. Resultados del modelo regresión lineal por componentes principales

Modelo	Coeficientes B	Error típico	T calculado	Significancia
(Constante)	8.9207	10.618	0.8402	0.4391
c1	47.668	27.825	1.7132	0.1474
c2	-22.88	19.172	-1.193	0.2863
c3	9.7288	15.757	0.6174	0.564

ANOVA					
Fuente	Suma de cuadrados	Grados de Libertad	Media cuadrática	F	Sig.
Regresión	561.9	3	187.3	0.9983	0.4655
Residual	938.1	5	187.62		
Total	1500	8			

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
0.612	0.3746	-6E-04	13.697

Ahora se aplica la técnica de regresión PLS en el paquete estadístico Minitab 15.0 para windows, para determinar el número de componentes se utiliza la técnica de la validación cruzada (crossvalidation), llegando a resumir las 9 variables independientes en un solo componente artificial, los resultados de la regresión PLS se dan en el siguiente cuadro.

Cuadro 4. Resultados del modelo regresión PLS

Modelo	Coeficientes B	Error típico	T calculado	Significancia
(Constante)	25.00	3.76	6.65	0.0003
t1	5.35	2.45	2.19	0.0649

ANOVA					
Fuente	Suma de cuadrados	Grados de Libertad	Media cuadrática	F	Sig.
Regresión	609.06	1	609.06	4.7853	0.0649
Residual	890.94	7	127.28		
Total	1500	8			

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
0.6372	0.406	0.3212	11.282

7. Conclusiones

En conclusión se observa que la regresión PLS brinda un mejor ajuste, y en comparación con la regresión por componentes principales la regresión PLS en este caso ha sintetizado la matriz X en una sola componente a diferencia de los tres componentes sintéticos de la regresión por componentes principales.

8. Bibliografía

[1] *Mardia, K.V.* (1997). "Análisis multivariante", Academic Press, London.

[2] *M. Barker.* (2003). "Partial least squares". Revista de Quimiometría, 17:166–173.

[3] *Vega Carmen* (2008). "Regresión por Mínimos Cuadrados Parciales con Aplicación en Regresión Logística". Tesis UMSA, Carrera de Estadística.

Índice

Un conejo estaba sentado delante de una cueva escribiendo, cuando aparece un zorro.

- Hola, conejo, que haces?

- Estoy redactando mi tesis doctoral sobre como los conejos comen zorros.

- Ja, ja, pero que dices?

- No te lo crees?. Anda, ven conmigo dentro de la cueva...

Entran los dos y al cabo de un ratito sale el conejo con la calavera del zorro y se pone a escribir. Al cabo de un rato llega un lobo.

- Hola, conejo, que haces?

- Estoy escribiendo mi tesis sobre como los conejos comen zorros y lobos.

- Ja, ja, que bueno, que chiste más divertido!

- Que no te lo crees?. Anda, ven dentro de la cueva, que te voy a enseñar algo!

Al cabo de un rato sale el conejo con la calavera de lobo, y empieza otra vez a escribir. Después llega un oso.

- Hola, conejo, que estás haciendo?

- Estoy acabando de escribir mi tesis sobre como los conejos comen zorros, lobos y osos.

- No te lo crees ni tú

- Bueno, a que no te metes en la cueva conmigo?

De nuevo se meten los dos en la cueva, y como era de esperar, un león enorme se tira encima del oso y se lo come. El conejo recoge la calavera del oso, sale fuera y acaba su tesis.

Moraleja: *Lo más importante no es el contenido de tu tesis, sino tu asesor.*