Comparación de métodos de regresión en la predicción de dióxido de carbono

Gladys Linares (Departamento de Investigaciones en Ciencias Agrícolas) / Adrián Saldaña (Departamento de Investigaciones en Ciencias Agrícolas) / Marcela Rivera (Facultad de Ciencias de la Computación) / Hortensia J. Cervantes (Facultad de Ciencias Físico-Matemáticas)
Benemérita Universidad Autónoma de Puebla
Av. San Claudio y 14 Sur, Puebla, México

Resumen: el objetivo de este estudio es comparar los métodos de regresión que se utilizan más frecuentemente en la práctica, a saber, regresión por mínimos cuadrados ordinarios, regresión paso a paso, regresión con componentes principales y regresión por mínimos cuadrados parciales. En este artículo se exponen los elementos teóricos de cada uno de los métodos mencionados y se aplican a un conjunto de datos reales en la predicción del dióxido de carbono en función de propiedades del suelo, en un período de seca, en el Parque Ecológico Coatzacoalcos, Veracruz, México.

Palabras claves: gases de efecto invernadero, multicolinealidad, regresión paso a paso, regresión con componentes principales, regresión por mínimos cuadrados parciales.

1. INTRODUCCIÓN

Los gases efecto invernadero, entre los que se encuentra el dióxido de carbono (CO_2), son de suma importancia debido a que ellos son continuamente emitidos y removidos en la atmósfera por procesos naturales sobre la tierra.

En la actualidad existe un gran interés científico en establecer la relación que guardan las emisiones de gases de efecto invernadero y las propiedades del suelo. En los problemas de predicción de gases efecto invernadero a partir de propiedades del suelo es bastante frecuente que las variables independientes del modelo sean altamente colineales, y por tanto, los estimadores de la regresión por mínimos cuadrados ordinarios (OLS, por sus siglas en inglés) no son adecuados porque poseen varianzas muy grandes.

Se han propuesto diferentes técnicas para manejar los problemas causados por la multicolinealidad. Entre ellos, puede citarse la regresión paso a paso (RPP), que es una técnica que permite reespecificar el modelo. Otro grupo de técnicas se encaminan a buscar estimadores sesgados de los coeficientes del modelo de regresión, pero con varianzas más pequeñas que los estimadores OLS. En este grupo se encuentran la regresión con componentes principales (RCP) y la regresión por mínimos cuadrados parciales (PLS, por sus siglas en inglés).

El propósito del presente trabajo es comparar estos cuatro métodos de regresión para seleccionar el mejor modelo que prediga las emisiones de dióxido de carbono en función de las propiedades del suelo, en un período de seca, en el Parque Jaguaroundi, Coatzacoalcos, Veracruz, México.

En la sección 2 se brindan las características esenciales de las regresiones OLS, RPP, RCP y PLS. En la sección 3 se aplican al problema antes mencionado y se comparan a través de diferentes estadísticos que muestran la capacidad predictiva de los mismos. Finalmente, en las secciones 4 y 5 se dan las conclusiones y se relacionan las referencias.

2. MÉTODOS DE REGRESIÓN: OLS, RPP, RCP Y PLS

En esta sección se explican los métodos OLS, RPP, RCP y PLS.

El modelo de regresión para estos métodos puede escribirse como:

$$y = 1\beta_o + X\beta + \varepsilon \qquad (1)$$

donde,

y es un vector n x 1 de observaciones de la variable dependiente, es una constante desconocida,

X es una matriz n x p que consiste de n observaciones de las p variables,

β_o es un vector p x 1 de coeficientes de la regresión (parámetros desconocidos),

β es un vector p x 1 de coeficientes de la regresión (parámetros desconocidos), y

ε es un vector n x 1 de errores independientes e idénticamente distribuidos con media cero y varianza σ^2.

Si las variables incluidas en la matriz X y en el vector y están centradas, la ecuación (1) puede escribirse de manera más simple como:

$$y = X\beta + \varepsilon \qquad (2)$$

2.1. Regresión mínimo cuadrática ordinaria (OLS)

Cuando la matriz X tiene rango completo p, el estimador OLS se obtiene minimizando la suma de cuadrados de los residuos. Este estimador es un vector de dimensión p x 1 cuya expresión es

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'\gamma \qquad (3)$$

y brinda estimadores insesgados de los elementos del vector de parámetros desconocidos del modelo con varianza mínima para alguna función lineal de las observaciones. Cuando las variables independientes están altamente correlacionadas, $X'X$ es mal condicionada y la varianza de los estimadores OLS se hace muy grande, y por tanto, son estimadores muy inestables.

2.2. Regresión paso a paso (RPP)

La RPP permite la re-especificación del modelo. Este procedimiento comienza con la hipótesis de que no hay regresores en el modelo además de la ordenada al origen. El primer regresor que se selecciona para entrar al modelo es el que tenga la máxima correlación simple con la variable respuesta, luego en cada paso se reevalúan todos los regresores que habían entrado antes al modelo, mediante sus estadísticas parciales F. Un regresor agregado en una etapa anterior puede volverse redundante, debido a las relaciones entre él y los regresores que ya están en la ecuación. Si la estadística F de una variable es menor que la F de salida, esa variable se elimina del modelo. En este método de regresión se requieren dos valores de corte, la F de entrada y la F de salida.

2.3. Regresión con componentes principales (RCP)

La RCP es una de las maneras de tratar los problemas de mal condicionamiento de matrices. Básicamente lo que hace es obtener el número de componentes principales que brinda la variación máxima de *X*. Realmente es un método de regresión lineal en que la respuesta es regresada sobre los componentes principales de la matriz *X*.

La expresión del estimador en RCP es

$$\beta_{RCP} = V_m \alpha_m \qquad (4)$$

donde V_m es una matriz que consiste de los primeros m vectores de norma unidad,

$$\alpha_m = (Z'_m Z_m)^{-1} Z'\gamma \qquad (5)$$

Z son las componentes principales y m es el número de estas componentes principales retenidas en el modelo. RCP da estimadores sesgados de los parámetros.

2.4. Regresión mínimo cuadrática parcial (Regresión PLS)

Otro método para construir modelos predictivos cuando las variables independientes son muchas y altamente colineales es el PLS. Para llevar a cabo la regresión de y con las variables independientes $X_1,\ldots,X_p$, PLS trata de encontrar nuevos factores que desempeñan el mismo pa-

pel que las *X*'s. Estos nuevos factores se llaman variables latentes o componentes. Análogamente a RPC, cada componente es una combinación lineal de X_1, ...,X_p, pero mientras RCP usa solo la variación de X para construir los nuevos factores, PLS usa tanto la variación de *X* como de y para construir los nuevos factores que se usarán como variables explicatorias del modelo.

Existen diferentes algoritmos para obtener los estimadores PLS, pero los más usados son el NIPALS y el SIMPLS.

3. ESTUDIO COMPARATIVO DE OLS, RCP Y PLS EN LA PREDICCIÓN DEL DIÓXIDO DE CARBONO

Antes de comparar la capacidad predictiva de los modelos es útil introducir varias medidas del ajuste del modelo a los datos y de la fuerza de predicción de esos modelos.

La Suma de Cuadrados de Error de Predicción, conocida como la estadística PRESS, se considera una medida de lo bien que funciona un modelo de regresión para predecir nuevos datos. Se define como la suma de los residuales PRESS al cuadrado que son los residuos que se obtienen entre el valor observado y el valor predicho de la i-ésima respuesta observada, basado en un ajuste de modelo con los n-1 puntos restantes de la muestra.

Con la estadística PRESS se puede calcular otro estadístico conocido como R^2 para la predicción, que se define como:

$$R^2_{\text{predicción}} = 1 - (PRESS/\text{Suma de Cuadrados Total}) \quad (6)$$

Una aplicación muy importante de estos estadísticos es comparar modelos de regresión. En general un modelo con pequeño valor de PRESS es preferible a uno con *PRESS* grande. El R^2 predicción se interpreta de manera similar al estadístico R^2 utilizado usualmente para medir la bondad del ajuste del modelo a los datos: a mayores valores de estos estadísticos mayor es la bondad del ajuste y mayor es la capacidad predictiva del modelo.

Dado que el uso principal de los modelos de regresión que utilizaremos para la predicción del dióxido de carbono es la predicción de futuras observaciones utilizaremos la estadística *PRESS* y el R^2 predicción para seleccionar el mejor modelo.

3.1 Problema de estudio

El Parque Ecológico Jaguaroundi, en Coatzacoalcos, Veracruz, México, se ubica dentro del complejo petroquímico denominado la Cangrejera, correspondiente a las instalaciones de Petróleos Mexicanos (PEMEX) y es una zona que presenta diferentes grados de perturbación debido a actividades de dos tipos. La primera es que esta zona fue usada como fuente de materiales de construcción para el complejo industrial y como lugar de confinamiento de materiales de desecho durante la construcción y la segunda es un pertinaz pastoreo de ganado bovino realizado por campesinos de los alrededores.

Las propiedades físicas y químicas de las 24 muestras de suelo tomadas se analizaron por el grupo de Edafología del Instituto de Geología de la Universidad Autónoma de México. En el período de seca se consideraron 16 predictores.

Los resultados que se presentan en cada método fueron obtenidos en MINITAB 15.

3.2 Modelos ajustados por todos los métodos de predicción

A continuación se muestran los resultados obtenidos por cada método de regresión en el problema bajo estudio.

3.2.1 Regresión mínimo cuadrática ordinaria (OLS)

La tabla 1 muestra los coeficientes del modelo de regresión OLS, los errores estándar de los coeficientes, el estadístico t de Student con sus correspondientes valores p y los factores de inflación de la varianza (VIF). Estos últimos indican los graves problemas de multicolinealidad ya que todos son mayores que 1. Aunque el valor de R^2 es 82.6 %, el R^2 predicción es 0 %. Existe evidencia de que este modelo no es adecuado por las razones antes expuestas.

3.2.2 Regresión paso a paso (RPP)

En este método de regresión se requieren dos valores de corte, la *F* de entrada y la *F* de salida, y hemos preferido definirlas como iguales.

Tabla 1. Regresión OLS

Predictores	Coef	SE Coef	T	P	VIF
Constante	3083	31741	0.97	0.363	
T Cámara °C	-17.70	12.34	-1.43	0.195	7.6
T Amb. °C	-30.76	47.83	-0.64	0.541	23.6
T Suelo °C	16.05	56.63	0.28	0.785	103.7
Humedad prom. %	5.18	4.44	1.17	0.281	20.0
Altitud msnm	-32.82	17.86	-1.84	0.109	36.6
Salinidad	-631.60	981.70	-0.64	0.540	14.7
N Total %	560.60	612.70	0.91	0.391	27.6
Carbono mg/g	-5.21	7.05	-0.74	0.484	28.4
P Mg/kg	170.80	124.00	1.38	0.211	7.3
pH	-74.70	271.40	-0.28	0.791	9.2
CE mS/cm	-1211.00	1144.00	-1.06	0.325	54.0
D. Ap. g/cc	-1.50	354.60	0	0.997	15.8
CMRA	3.25	2.56	1.27	0.245	3.9
Respiración basal	-0.49	0.24	-2.03	0.082	3.2
%arcilla	-22.01	30.59	-0.72	0.495	80.3
%arena	-7.66	15.68	-0.49	0.640	46.0

Se tomó, en todos los casos, una tasa α de error tipo I igual a 0.15, para generar *F* de entrada y *F* de salida.

La tabla 2 muestra los dos pasos llevados a cabo por el procedimiento. Obviamente el mejor modelo es el de dos predictores, a saber, altitud y porciento de nitrógeno total, ambos significativos al 5 %. La bondad del ajuste que muestra el R^2 es del 44.14 % mientras que la capacidad predictiva mostrada por el R^2 de predicción es sólo del 19.07 %.

Tabla 2. Regresión paso a paso

Paso	1	2
Constante	439.8	279.3
Altitud msnm	-10.5	-7.7
T-Value	-3.17	-2.34
P-Value	0.004	0.029
N Total %		287
T-Value		2.19
P-Value		0.04
S	139	129
R-Sq	31.33	44.14
PRESS	554306	502437
R-Sq (pred)	10.71	19.07

3.2.3 Regresión con componentes principales

El análisis de componentes principales se realizó sobre la matriz de correlaciones (tabla 3). Puede apreciarse que 7 componentes principales explican el fenómeno en casi su totalidad (93.7 %). En la primera componente se destacan como variables muy importantes el carbono, el nitrógeno total, la conductividad eléctrica (CE) y la densidad aparente (D.Ap). En esta componente se muestra oposición entre el carbono, la conductividad eléctrica y el nitrógeno (signos negativos) y la densidad aparente (signo positivo).

Por simplicidad se decidió realizar la regresión lineal simple con la primera componente, dado que a pesar de mostrar un R^2 de 34.7 %, la prueba F obtenida en el análisis de varianza que se muestra en la tabla 3(C) resultó significativa. El R^2 predicción es 28.2 %.

La ecuación de regresión obtenida es:

$$CO_2 = 124 - 41.1\ (CP1).$$

Tabla 3. Análisis de componentes principales de la matriz de correlación

(A) Valores propios

Valores propios	5.5589	3.3737	2.2688	1.5373	1.1497	0.6564	0.4496
Proporción	0.347	0.211	0.142	0.096	0.072	0.041	0.028
Prop acum	0.347	0.558	0.700	0.796	0.868	0.909	0.937

(B) Vectores propios

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
T Cámara °C	-0.04	0.236	-0.261	0.424	0.367	-0.412	-0.426
T Amb. °C	-0.14	0.377	0.31	-0.229	-0.106	-0.008	0.101
T Suelo °C	-0.068	0.507	0.037	-0.232	0.053	-0.088	-0.026
Humedad prom. %	-0.278	0.174	0.093	0	-0.489	-0.499	0.059
Altitud msnm	0.275	0.332	-0.076	-0.253	-0.104	0.257	0.01
Salinidad	0.07	-0.471	0.154	0.142	-0.165	-0.174	0.111
N Total %	-0.341	0.007	-0.275	-0.125	-0.23	-0.076	0.062
Carbono mg/g	-0.375	0.003	-0.259	-0.032	-0.041	0.021	-0.158
P Mg/kg	0.133	-0.242	-0.372	-0.463	-0.073	0.007	-0.005
pH	-0.168	0.024	0.48	0.25	-0.263	0.218	-0.2
CE mS/cm	-0.369	-0.114	-0.253	-0.089	-0.04	0.019	0.161
D. Ap. g/cc	0.354	-0.039	0.105	0.004	-0.211	-0.022	-0.188
CMRA	-0.304	-0.157	0.034	-0.136	-0.064	0.427	-0.661
Respiración basal	0.27	-0.094	0.072	-0.376	-0.17	-0.454	-0.468
% arcilla	-0.25	-0.27	0.319	-0.191	0.18	-0.281	0.07
% arena	0.154	0.054	-0.319	0.362	-0.577	0.049	-0.006

Tabla 3 (C). Análisis de varianza para CO_2 por RCP (una componente)

Fuente variación	gl	SC	CM	F	P
Regresión	1	215710	215710	11.71	0.002
Error	22	405092	18413		
Total	23	620802			

3.2.4 Regresión mínimo cuadrática parcial (Regresión PLS)

La tabla 4 muestra los resultados de la regresión por mínimos cuadrados parciales. El número de componentes fue seleccionado por validación cruzada. Puede apreciarse que el modelo con una sola componente tiene la mayor capacidad predictiva y el Análisis de varianza mostró que la prueba *F* es significativa.

Los coeficientes estandarizados más altos con signos positivos aparecen en los predictores altitud, nitrógeno total y conductividad eléctrica. La densidad aparente tiene el coeficiente negativo más alto.

La tabla 5 muestra la comparación entre los modelos según su capacidad predictiva. Puede observarse que el modelo con mayor capacidad predictiva para estos datos, correspondientes al período de secas, es la RCP que tiene una estadística PRESS de 446335 más baja que los restantes métodos y R^2 predicción igual a 28.1 %, más alto que el PLS en más de 23 unidades porcentuales y superando el modelo de regresión paso a paso en 9 unidades porcentuales.

Tabla 4 (A). Regresión PLS. Selección de modelo y validación para CO_2 por PLS

Componentes	X Variance	Error SS	R-Sq	PRESS	R-Sq (pred)
1	0.366301	367409	0.408171	590451	0.0488898
2		323969	0.478144	853490	0.0000000
3		279968	0.549022	961751	0.0000000
4		192443	0.690008	1334971	0.0000000
5		176083	0.716363	1422984	0.0000000
6		153515	0.752715	1448637	0.0000000
7		142103	0.771097	1532742	0.0000000
8		122791	0.802205	1690775	0.0000000
9		118290	0.809457	1697396	0.0000000
10		114700	0.815238	1661211	0.0000000
11		112521	0.818749	1844719	0.0000000
12		111787	0.819932	1977837	0.0000000
13		110306	0.822318	2431989	0.0000000
14		108772	0.824788	2522072	0.0000000
15		108099	0.825872	2684584	0.0000000
16		107791	0.826367	2679756	0.0000000

Tabla 4 (B). Regresión PLS. Análisis de Varianza para CO_2 por PLS (una componente)

Fuente variación	gl	SC	CM	F	P
Regresión	1	253393	253393	15.17	0.001
Error	22	367409	16700		
Total	23	620802			

Tabla 4 (C). Regresión PLS. Coeficientes de regresión

	CO_2	CO_2 Estandarizado
Constant	2.9157	0.0000000
T Cámara °C	-0.6885	-0.0242192
T Amb. °C	0.1285	0.0020571
T Suelo °C	-0.6667	-0.0188711
Humedad prom. %	0.5007	0.0794491
Altitud msnm	-1.7687	-0.0943516
Salinidad	24.616	0.0151164
N Total %	67.9816	0.0917254
Carbono mg/g	0.6621	0.0788305
P Mg/kg	-1.2301	-0.0042329
pH	22.8037	0.0401527
CE mS/cm	90.7248	0.0917562
D. Ap. g/cc	-40.4046	-0.0713172
CMRA	0.7163	0.0874503
Respiración basal	-0.0569	-0.0660981
% arcilla	1.4337	0.0661290
% arena	-0.0946	-0.0064409

Tabla 5. Comparación de 4 modelos de regresión para predecir CO_2

	Regresión OLS	Regresión paso a paso	RCP (1 CP)	PLS (1 CP)
PRESS	2680053	502437	446335	590451
R^2 predicción	0.00 %	19.07 %	28.10 %	4.89 %

4. CONCLUSIONES

Se ha llevado a cabo un estudio comparativo entre los métodos de regresión OLS, RPP, RCP y PLS en el estudio de predicción de CO_2 en función de propiedades del suelo con datos que presentaban multicolinealidad severa. La comparación se realizó bajo el criterio de mejor capacidad predictiva. La RCP con una componente mostró la estadística PRESS más baja y el R^2 predicción más alto entre los cuatro modelos considerados.

5. REFERENCIAS

[1] Montgomery, D. C., Peck, E. A. y Vining, G. G., Introducción al Análisis de Regresión Lineal, Compañía Editorial Continental. México.

[2] Saldaña, J.A., Ruiz-Suárez, L.G., Hernández, J.M. y Morales. B.M. (2006). Emisiones de gases efecto invernadero en suelos perturbados con diferente cobertura vegetal en Coatzacoalcos, Veracruz, México. (por aparecer).

[3] Yeray, O y Goktas, A., A Comparison of Partial Least Squares Regression with other Prediction Methods, Hacettepe Journal of Mathematics and Statistics. 31, 99-111 (2002).on-info.html, 1999.