# EL USO DE RECOCIDO SIMULADO PARA LA SELECCION DEL MEJOR MODELO DE REGRESION: EL CASO $L_2$

Sira M. Allende Alonso*, Carlos N. Bouza Herrera y Gemayqzel Bouza Allende, Universidad de La Habana

**RESUMEN**

En muchos problemas el ajuste de un modelo de regresión es requerido. Uno de los problemas a resolver para hacer uso efectivo de criterios de ajuste de la regresión es el poder seleccionar modelos. Comúnmente esto se hace al usar el criterio de los Mínimos Cuadrados dependiendo de pruebas basadas en la normalidad de los errores. En este trabajo se sugiere solucionar esto a partir de heurísticas basadas en el Recocido Simulado. Estas buscan la disminución de la suma de los cuadrados la que usan como función objetivo. Se desarrollan algoritmos y se hace una evaluación de las propuestas analizando ejemplos clásicos que aparecen analizadas en libros de texto. El comportamiento del método propuesto y algoritmizado aparece como adecuada pues se obtienen los mejores ajustes reportados.

Palabras clave: regresión lineal, modelo óptimo, optimización paramétrica.

**ABSTRACT**

In many problems is needed to fit a regression model. For using effectively different adjustment criteria one of the problems to be solved is how to select the best regression model. Commonly when the method of the Least Squares is used assuming the normallity of the errors. In this paper we suggest to solve this problem by using a Simulated Annealling based heuristics. The method look for the disminution of the residual sum of squares using it as objective function. Algorithms are developed and an evaluation of the proposals is made by analyzing classic examples from text books. The behavior of them seems to be adequate because they identify the best fitted models.

Key words: linear regression, optimal model, parametric optimization.

MSC: 62J05

## 1. PLANTEAMIENTO DEL PROBLEMA

El modelo de regresión lineal usual está dado por

$$Y_i = x_i\beta + \varepsilon_i$$

donde

$$\beta \in \Re^{p+1}$$

$$x_i = (1, x_{i1}, \ldots x_{ip})$$

y

$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$$

Generalmente se supone que la distribución de los errores $\varepsilon_i$ 's es una normal y que son independientes. Este es considerado como un modelo central F*. Bajo estas hipótesis el método de los Mínimos Cuadrados (MC) es óptimo. Cuando observamos valores poco probables de los errores dudamos de que F* los genere. Tal es el caso en muchos problemas económicos en los que factores externos afectan las respuestas de los agentes y los sistemas. Decimos que un error es extremal si $\varepsilon \notin [-3\sigma, 3\sigma]$. Entonces se considera que en realidad estas variables tienen una distribución no normal y que el modelo conque trabajamos es una distribución contaminada de F* con otra F'. Entonces los errores tienen una distribución que pertenece a una vecindad $\lambda$-contaminada

*sira@matcom.uh.cu

$$\wp = \{F | F = \lambda F^* + (1 - \lambda)F', \lambda \in \{0,1]\}$$

La Estadística Robusta sugiere el uso de un M-estimador del tipo un-paso (one-step M-Estimator), Bickel (1984). Estas son funciones del tipo:

$$B(n, Y_1,..., Y_n) = \beta^*(Y_1,..., Y_n) + \sigma_0 \sum \Psi\left(\frac{Y_i - x_i^T \beta^*(Y_1,..., Y_n)}{\sigma_0}\right)$$

donde $\beta^*(Y_1,..., Y_n)$ es un estimador inicial de $\beta$, generalmente obtenido mediante el método de los MC, y $\sigma_0$ es un estimador de la escala. Las condiciones para su normalidad asintótica requiere de una serie de hipótesis sobre la regularidad de la función de distribución, ver Müller (1994).

Dodge (1984) propuso el siguiente problema de programación paramétrica para resolver el ajuste de una $F \in \wp$:

$$P1: \text{Min}_\beta\left(\lambda \sum_{i+1}^{n} \delta^2{}_i + (1-\lambda) \sum_{i+1}^{n} |\delta_i|\right)$$

sujeto a:

$$X\beta - \delta \leq Y$$

$$-X\beta - \delta \leq -Y$$

$$\lambda \in [0,1] \qquad \text{fijo}$$

$$\delta = (\delta_1,..., \delta_n)^T, \delta_i = Y_i - x_i^T \beta$$

$$X = (x_1,..., x_n)^T$$

$$Y = (Y_1,..., Y_n)^T$$

Allende-Bouza (1993) consideraron el caso en que $\lambda$ era desconocido. Esto hace necesario determinar el parámetro de contaminación dentro del programa. El nuevo programa es:

$$P2: \text{Min}_{(\beta,\lambda)}\left(\sum_{i+1}^{n} \delta^2{}_i + \theta \sum_{i+1}^{n} |\delta_i|\right)$$

sujeto a

$$X\beta - \delta \leq Y$$

$$-X\beta - \delta \leq -Y$$

$$\theta = \frac{\lambda}{1-\lambda}, \qquad \lambda \in [0,1[$$

$$\delta = (\delta_1,..., \delta_n)^T, \delta_i = Y_i - x_i^T \beta$$

$$X = (x_1,..., x_n)^T$$

$$Y = (Y_1,..., Y_n)^T$$

En ese trabajo se abordó el estudio del problema de Complementariedad Lineal asociado y se propuso un algoritmo para resolverle. Este fue mejorado en Suárez Fariñas **et al**. (1997). Este reportó la posibilidad de resolver el problema brindando una partición del espacio paramétrico

$$\cup_{k=1}^{K} \Lambda_k=[0,1]$$

junto con el vector de coeficientes óptimos de la regresión $\beta(\lambda | \lambda \in \Lambda_k)$ para cada subintervalo. Queda en manos del decisor el seleccionar el mejor modelo. Es lógico que este use el modelo en el que $\lambda^*$ haga mínima la función objetivo.

Un problema que debe ser resuelto para todos estos modelos alternativos es el de seleccionar las variables que deben entrar en el modelo. A partir de tener p variables $X_1,...,X_p$ el decisor desea evaluar otro modelo con p' variables. Esto lo hace al definir un conjunto de transformaciones de las variables iniciales y evaluar si estas deben entrar a formar parte del modelo. Eventualmente, incluirles puede conllevar la eliminación de una de las variables analizadas. Tomemos $M(G_1,...,G_m)$ como una familia de tales transformaciones. Se analizará para una variable $X_j$ si la inclusión de $G_h(X_j)$ redunda en una mejor ecuación de regresión. Por ejemplo podemos usar $M(X_j^2 , \log X_j, 0)$. Al aceptar la transformación 0 lo que se hace es eliminar $X_j$ en el nuevo modelo ajustado.

Los criterios básicos utilizados en el estudio del ajuste de los modelos son el Ascendente y el Descendente, ver Johnson y Wichern (1998). En el primero se parte de un grupo exiguo de variables, generalmente una, y se va incrementando su número. En el segundo se utiliza un modelo con muchas variables, usualmente todas, y se van eliminando las menos importantes en la predicción. El método Paso a Paso evalúa en cada modelo aceptado si este es mejorado con algunas de las variables eliminadas previamente. Este es más caro desde el punto de vista computacional pero más confiable. Bajo las hipótesis para las cuales el método de los MC es óptimo la entrada o eliminación de variables se lleva a cabo usando pruebas del tipo F.

En este trabajo se sugiere solucionar esto a partir de heurísticas basadas en el Recocido Simulado. Estas buscan la disminución de la suma de los cuadrados la que usan como función objetivo. El procedimiento es eminentemente no paramétrico pues no se basa en ninguna hipótesis sobre la distribución de los errores. Se proponen y programan algoritmos y se hace una evaluación de las propuestas analizando ejemplos clásicos que parecen en la literatura. El comportamiento de la propuesta es valorada de adecuada pues se obtienen los mejores ajustes reportados con una alta frecuencia.

## 2. LOS ALGORITMOS

El método del Recocido Simulado (RS) es una poderosa herramienta de búsqueda estocástica que se ha hecho muy popular dado el amplio espectro de problemas que puede resolver. En particular en el área de la optimización combinatoria se ha hecho de un nicho investigativo.

El parámetro m es llamado dimensión del espacio y se define una función de costo $C:S \longrightarrow \Re$ que le asigna un valor real a cada configuración. El objetivo es hallar la configuración óptima: $z^* \in S$ tal que $\forall z' \in S$ entonces $C(z^*) \leq C(z')$. La convergencia al óptimo global se obtiene a partir de la modelación del proceso del Recocido Simulado mediante Cadenas de Markov, ver Pflug (1996).

En nuestro caso deseamos escoger cuál es la combinación de funciones $(G_1,..., G_m\}$ que mejor describe las relaciones descritas por una regresión para un determinado conjunto de datos. Tenemos que m es el total de funciones, sea $S = \{z \mid z = (z_1,...,z_m) \in \{0,1\}^m\}$ el espacio de todas las combinaciones posibles donde

$$z_i = \begin{cases} 1 & \text{si } G_i \text{ entra en la ecuación} \\ 0 & \text{en otro caso} \end{cases}$$

El algoritmo consiste en, partiendo de una combinación, pasar a un vecino el cual se obtiene agregando o quitando una y solo una transformación. Si se obtiene una mejora de C(z) se acepta. Peores configuraciones

se aceptarán solo con una cierta probabilidad menor que uno. Esto garantiza la convergencia a un mínimo global evitando que el algoritmo termine en una solución local, ver Pflug (1996). En nuestro caso la función de costo estará definida como el error cuadrático cometido por aproximar el conjunto de datos y por las funciones presentes en z.

La variante que usaremos consiste en:

1. Introducir los datos correspondientes a Y y a X y las transformaciones ($G_1$,...,$G_m$}. Escoger una configuración inicial, $z_0$, y hallar C(z). Fijar la probabilidad (p) conque se permitirá entrar una función en la combinación y con cual salir (q).

2. $C^* = C(z_0)$

3. While t > e do
   Repeat for r = 1 to R
   while h < H
   $z = z_r$
   Seleccionar aleatoriamente $u \in (0,1,...,m\}$
   Generar un número aleatorio r.
   Si $z_u = 1$ y r<p respectivamente (si $z_u = 0$ y r < q)
   agregar (quitar) $G_u[x_u]$, i = 1,...,n.
   resolver el nuevo problema con $z = (z_1,z_2,,...z_{u-1}, 1-z_u , z_{u+1},....z_m)$
   End If
   Si (C(z) > C* y exp([C(z)-C*]/t) > rand[0,1]) o si $C(z) \leq C^*$
   $z_{r+1} = z$
   h = H
   End If
   h = h+1
   End while
   r = r+1
   End repetir
   $t = \alpha t$

end while
end

Cabe destacar que este algoritmo incluye las conocidas estrategias de selección de modelos de regresión denominadas Ascendente, Descendente y Paso a Paso al fijar convenientemente los parámetros p y q. Note que si p = 1 y q = 0 solo se agregarán funciones a la configuración y se implementa el método Ascendente; p = 0, q = 1 implementa la estrategia Descendente y p = q = 1 la Paso a Paso. Cabe destacar que el usuario puede también escoger probabilidades entre 0 y 1, lo que modelaría toda una serie de familia de métodos de selección de modelos no estudiados en la literatura estadística usual.

## 3. ANALISIS DE LA PROPUESTA: EXPERIMENTOS CON PROBLEMAS CONOCIDOS

El comportamiento del algoritmo propuesto fue evaluado al hacer un estudio de problemas clásicos. Fue aplicado en 100 corridas independientes. El número promedio de iteraciones requeridas para obtener el modelo final, que se considera óptimo fue computada y el porciento de veces que cada uno de ellos fue el propuesto. El valor del coeficiente de determinación es brindado en las tablas. Como valores de r fueron usados 5 y 2. Las combinaciones de los valores de los parámetros p y q fueron (1,0), equivalente al criterio de búsqueda ascendente, (0,1), asociado al criterio descendente y (0,0) que identifica el Paso a Paso. Como el término independiente se asocia a una variable k es el número de variable explicativas más uno. El índice de la primera variable es el cero ($X_0 = 1$).

**Problema 1**. Datos de los ejercicios A,B y C del Capítulo 6, Draper-Smith (1980), n = 17 y k = 5.

**Tabla P1.1**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 5.

| Parámetros | Indice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 1,2 | 2,0 | 0,49 | 1 |
| | 1,2,4 | 4,7 | 0,63 | 1 |
| | 1,2,3,4, | 3,9 | 0,77 | 98 |
| (0,1) | 1,2,4 | 2,8 | 0,63 | 3 |
| | 1,2,3,4 | 1,6 | 0,77 | 97 |
| (0,0) | 1,2,3,4 | 61,4 | 0,77 | 100 |

Note en la Tabla P1.1 que el mejor modelo es el que contiene todas las variables y este es seleccionado con la mayor frecuencia.

**Tabla P1.2**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 2.

| Parámetros | Indice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 1,2 | 1,8 | 0,49 | 1 |
| | 1,2,4 | 3,0 | 0,63 | 2 |
| | 1,2,3,4, | 3,6 | 0,77 | 97 |
| (0,1) | 1,2,4 | 2,2 | 0,63 | 5 |
| | 1,2,3,4 | 1 | 0,77 | 95 |
| (0,0) | 1,2,3,4 | 55,3 | 0,77 | 100 |

Los resultados de la Tabla P1.2 son similares aunque la frecuencia conque se acepta el mejor modelo es ligeramente menor. Sin embargo hay mayor rapidez en determinar el modelo considerado óptimo.

Note que es un caso en el que el uso de otro criterio de optimización como el basado en la norma $L_1$ puede ser recomendado para ganar en precisión del ajuste.

**Problema 2**. Datos dados en Brownlee (1965), página 454 con n = 21 y k = 4.

**Tabla P2.1**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 5.

| Parámetros | Índice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 1 | 1,5 | 0,85 | 24 |
| | 1,2 | 6,1 | 0,91 | 76 |
| (0,1) | 1 | 8,7 | 0,85 | 5 |
| | 1,2 | 6,1 | 0,91 | 10 |
| | 1,2,3 | 7,9 | 0,91 | 85 |
| (0,0) | 1,2, | 58,2 | 0,91 | 93 |
| | 1,2,3 | 63,1 | 0,91 | 7 |

En este problema el uso de todas las variables o el excluir la tercera genera una diferencia entre los correspondientes valores de $R^2$ del orden de 0,04. Lo esperado es que el decisor use el modelo con menos variables. El criterio ascendente tiende a no incluir $X_2$ a pesar de la ganancia, el descendente a no eliminar $X_3$. Esta última tendencia es menor para el Paso a Paso. Para r = 5 esto es menos evidente. En todos los casos se elige con mucha frecuencia un modelo con $R^2 = 0.91$.

**Tabla P2.2**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 2.

| Parámetros | Índice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 1 | 1 | 0,85 | 33 |
| | 1,2 | 3,7 | 0,91 | 67 |
| (0,1) | 1 | 4,3 | 0,85 | 20 |
| | 1,2 | 3,6 | 0,91 | 20 |
| | 1,2,3,4 | 3,9 | 0,91 | 60 |
| (0,0) | 1,2, | 55,2 | 0,91 | 93 |
| | 1,2,3,4 | 59,2 | 0,91 | 7 |

**Problema 3**. Datos analizados en el Capítulo 7 del Draper-Smith (1980) con n = 10 y k = 5.

**Tabla P3.1**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 5.

| Parámetros | Índice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 4 | 1,6 | 0,89 | 25 |
| | 1,3,4 | 2,5 | 0,94 | 65 |
| | 1,2,3,4 | 4,2 | 0,97 | 10 |
| (0,1) | 1,2,4 | 4,9 | 0,89 | 20 |
| | 1,3,4 | 7,9 | 0,94 | 15 |
| | 1,2,3,4 | 1,6 | 0,97 | 65 |
| (0,0) | 4 | 78,3 | 0,89 | 5 |
| | 1,3,4 | 55,2 | 0,94 | 85 |
| | 1,2,3,4 | 65,4 | 0,97 | 10 |

Cuando r = 2 el criterio descendente no considera como un modelo posible el que incluye las variables $X_1$, $X_3$ y $X_4$ que es mejor que el que lo hace con $X_1$, $X_2$ y $X_4$ . Esto no ocurre con r = 5 aunque lo selecciona con menor frecuencia que el asociado a $X_1$, $X_3$ y $X_4$.

**Tabla P3.2**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 2.

| Parámetros | Índice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 4 | 1 | 0,89 | 25 |
| | 1,3,4 | 2 | 0,94 | 65 |
| | 1,2,3,4 | 3,4 | 0,97 | 10 |
| (0,1) | 1,2,4 | 3,3 | 0,89 | 33 |
| | 1,2,3,4 | 2,6 | 0,97 | 67 |
| (0,0) | 4 | 66,3 | 0,89 | 5 |
| | 1,3,4 | 59,2 | 0,94 | 15 |
| | 1,2,3,4 | 70,4 | 0,97 | 750 |

**Problema 4**. Datos de la página 352 de Draper-Smith (1980) con n = 25 y k = 11.

**Tabla P4.1**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 5.

| Parámetros | Indice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 6,8 | 2,6 | 0,85 | 100 |
| (0,1) | 8 | 4,4 | 0,71 | 5 |
| | 6,8 | 2,5 | 0,85 | 95 |
| (0,0) | 6,8 | 59,7 | 0,85 | 100 |

En este problema la optimalidad del modelo que usa las variables $X_6$ y $X_8$ es evidente. La ganancia con el uso de las otras variables es despreciable. Los cambios en r no son significativos.

**Tabla P4.2**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 5.

| Parámetros | Índice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 6,8 | 1 | 0,85 | 100 |
| (0,1) | 8 | 3,2 | 0,71 | 20 |
| | 6,8 | 2,9 | 0,85 | 80 |
| (0,0) | 6,8 | 44,8 | 0,85 | 100 |

**Problema 5**. Datos de Hald (1952), página 647 con n = 13 y k = 5.

**Tabla P5.1**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 5.

| Parámetros | Índice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 1,2 | 2,9 | 0,98 | 70 |
| | 1,3,4 | 2,3 | 0,98 | 5 |
| | 1,2,3 | 3,1 | 0,98 | 25 |
| (0,1) | 1,2,3 | 1,6 | 0,98 | 50 |
| | 1,3,4 | 1,3 | 0,98 | 45 |
| | 2,3,4 | 2,2 | 0,98 | 5 |
| (0,0) | 1,2 | 49,3 | 0,98 | 10 |
| | 1,2,3 | 51,1 | 0,98 | 45 |
| | 1,3,4 | 48,7 | 0,98 | 35 |
| | 2,3,4 | 47,2 | 0,97 | 5 |
| | 1,2,3,4 | 28,2 | 0,98 | 5 |

En este problema varios modelos poseen la misma valoración aproximada de $R^2$. Las diferencias son del orden de las milésimas. Para r = 5 el método ascendente tiende a preferir con mayor frecuencia el modelo que solo usa a $X_1$ y $X_2$, el que difiere del que usa también a $X_3$ solo en -0,03. Esto hace aceptable el no seguir buscando otra variable. Este también sugiere usar el que usa las variables 1, 3 y 4 solamente que cuyo resultado solo empeora el asociado a las primeras tres variables en -0,09. El descendente tiende a preferir no eliminar y el Paso a Paso es el menos conservador incluyendo una mayor diversidad de modelos. Esta tendencia nuevamente es más marcada al usar r = 2.

**Tabla P5.2**. Selección de modelos usando búsqueda mediante Recocido Simulado con r = 2.

| Parámetros | Índice de las variables en el modelo | Promedio de las iteraciones | $R^2$ | % de veces |
|---|---|---|---|---|
| (1,0) | 1,2 | 2,2 | 0,98 | 60 |
| | 1,3,4 | 2,1 | 0,98 | 35 |
| | 1,2,3 | 2,4 | 0,98 | 5 |
| (0,1) | 1,2,3 | 1,3 | 0,98 | 50 |
| | 1,3,4 | 1,3 | 0,98 | 30 |
| | 2,3,4 | 1,8 | 0,98 | 20 |
| (0,0) | 1,2 | 49,3 | 0,98 | 30 |
| | 1,2,3 | 51,1 | 0,98 | 30 |
| | 1,3,4 | 48,7 | 0,98 | 15 |
| | 2,3,4 | 47,2 | 0,97 | 5 |
| | 1,2,3,4 | 28,2 | 0,98 | 20 |

Estos experimentos sugieren que los criterios identifican el modelo óptimo con una frecuencia mucho mayor que los alternativos. Cada uno de ellos tiende a conservar el modelo previo si no hay un aumento significativo en $R^2$. El método Paso a Paso realiza más iteraciones lo que le hace considerablemente más lento. El efecto de r es significativo pues este acentúa las características conservadoras de los criterios Ascendente y Descendente.

**AGRADECIMIENTOS**

Parte de este trabajo se desarrolló durante una visita del autor a la Universidad A Coruña.

**RECONOCIMIENTOS**

Este trabajo se llevó a cabo soportado parcialmente por una Beca DAAD para uno de los autores. Sus resultados se inscriben en los objetivos del Proyecto Alma Máter Modelos óptimos para aplicaciones de la estadística a problemas médicos y de la Educación Superior.

**REFERENCIAS**

ALLENDE, S. and C. BOUZA (1993): A parametric programming approach to the estimation of the coefficient of a linear regression model. En “Approximation and Optimization”, P. Lang Verlag, Berlin, 9-21.

BROWNLEE, K.A. (1965): **Statistical Theory and Methodology in Science and Engineering**, J. Wiley, New York.

DODGE, Y. (1984): Robust estimation of the regression coefficients by minimising a convex combination of LS and LAD. **Comp. Stat. Quart. J**. 1, 139-153.

DRAPER, N.H. and H. SMITH (1980): **Applied Regression Analysis**. Ed. Pueblo y Educación, Habana.

HALD, A. (1952): **Statistical Theory with Engineering Applications**, J. Wiley, New York.

JOHNSON, R.A. and D.W. WICHERN (1998): **Applied Multivariate Analysis**, Prentice Hall, New Jersey.

MÜLLER, CH.H. (1994): “Optimal designs for robust estimation in conditionally contaminated models”, **J. of Stat. Planning and Inference**, 38, 125-140.

PFLUG, G.CH. (1996): Optimization of Stochastic Models: The interface between Simulation and Optimization, Kluwer Academic Publishers, Massachussetts.

SUAREZ FARIÑAS, M., S.A. GARCET RODRIGUEZ, S. ALLENDE ALONSO and C. BOUZA HERRERA (1997): “El problema de complementariedad lineal: algoritmos de solución y aplicación al ajuste del modelo de regresión lineal multivariado con LS-LAD”, **Inv. Operacional**, 18, 11-56.