NOTA TÉCNICA

EFECTO DEL MÉTODO DE SELECCIÓN DE LA MUESTRA EN LA ESTIMACIÓN DE LOS PARÁMETROS EN MODELOS DE REGRESIÓN LINEAL

Rafael Calama Sainz

Dpto. Selvicultura y Gestión Forestal. CIFOR-INIA. Ctra. de la Coruña km 7,5. 28040-MADRID (España). Correo electrónico: rcalama@inia.es

Resumen

El efecto que el método de selección muestral pueda tener sobre la estimación de los parámetros en los modelos de regresión lineal es un aspecto poco estudiado en la literatura estadística. En el presente trabajo se aplican métodos de extracción de muestras mediante técnicas de Monte Carlo sobre una población simulada al objeto de evaluar el efecto que tres tipos de selección (1) aletoria; (2) dirigida y (3) proporcional a la variable explicativa tienen sobre la precisión y sesgo de los estimadores de los parámetros. Se comprueba que la selección aleatoria es la que conduce a estimaciones de los parámetros con menor sesgo, mientras que la simulación dirigida permite obtener los estimadores más precisos. El presente trabajo abre una línea para la identificación de la muestra óptima (tamaño y método de selección) para construcción de modelos forestales.

Palabras clave: *Estimador, Simulación, Modelo de regresión, Monte Carlo, Muestra*

INTRODUCCIÓN

El modelo de regresión lineal es una técnica cuantitativa de análisis estadístico que permite relacionar una variable de respuesta de tipo continuo y con una o varias variables explicativas xi también de tipo continuo, mediante la siguiente función lineal:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e = \beta_0 + \Sigma\beta_i x_i + e = \hat{y} + e$$

Donde β_i representa los parámetros del modelo de regresión, $\hat{y}$ el valor predicho para la observación de la variable de respuesta *y*; *e* el término del error independiente, que define la diferencia entre el valor observado y el valor predicho. El objeto del análisis de regresión es determinar el valor de los parámetros β_i que mejor expliquen la relación entre *y* y x_i. Como habitualmente tendremos observaciones procedentes de una muestra de la población (y no de toda la población en conjunto), no obtendremos el valor real de estos parámetros β_i, sino un estimador de los mismos, que denomináramos b_i. El método de estimación de parámetros más habitual es el de mínimos cuadrados ordinarios, basado en calcular los parámetros que minimizan la suma de cuadrados de los términos *e* sobre el total de las observaciones de la muestra. Este método permite obtener los estimadores más eficientes (insesgados y de menor varianza) de los parámetros poblacionales, siempre y cuando se verifiquen los supuestos básicos de

ISSN: 1575-2410

normalidad, homocedasticidad en la varianza e independencia de los términos del error *e*. En el caso de incumplimiento de los supuestos, el método debe sustituirse por otros métodos como la estimación mediante máxima verosimilitud y la regresión por mínimos cuadrados generalizados o ponderados (SEARLE, 1971).

Sin embargo, e incluso en el caso de verificación de los supuestos básicos de regresión, existe otro aspecto adicional de interés, muy pocas veces evaluado, como es el impacto que sobre la estimación de los parámetros pueden tener tanto el tamaño de la muestra como el método de selección de la misma dentro de la población. Consideremos el caso más sencillo, como es el del modelo de regresión lineal simple (una única variable explicativa x) que podemos expresar como:

$$y = \beta_0 + \beta_1 x + e$$

Para este modelo, el valor esperado de los estimadores muestrales de mínimos cuadrados ordinarios para β_0 y β_1, que denominaremos b_0 y b_1, viene dado por las siguientes expresiones (ecs. 1-2, ver p.ej. MYERS, 1990: 13):

$$b_0 = (\bar{y} - b_1\bar{x}) \qquad [1]$$

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad [2]$$

Los estimadores muestrales b_0 y b_1 llevan asociados a su vez un nivel de incertidumbre, definido por el valor de sus estimadores de varianzas $S^2(b_0)$ y $S^2(b_1)$ (ecs. 3-4, MYERS, 1990: 14-15):

$$S^2(b_0) = S^2(e)\left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right] \qquad [3]$$

$$S^2(b_1) = \frac{S^2(e)}{\sum_i (x_i - \bar{x})^2} \qquad [4]$$

Donde $\bar{x}$ e $\bar{y}$ representan el valor medio de las observaciones muestrales para la variable explicativa x y la variable dependiente *y*; x_i e y_i representan el valor observado para la variable explicativa x y la variable dependiente *y* en la i-ésima observación de la muestra; *n* es el número de observaciones (tamaño) de la muestra, y $S^2(e)$ representa la varianza del error, supuesta constante para todo el rango de y (condición de homocedasticidad). De acuerdo a las expresiones anteriores, es fácil comprobar que tanto el valor del estimador muestral de los parámetros como el de sus varianzas asociadas dependen de la varianza del error (constante en el caso de verificación de la condición de homocedasticidad), de las medias muestrales $\bar{x}$ e $\bar{y}$, del tamaño de muestra *n*, y de manera principal de la dispersión de los valores muestrales de la variable x alrededor de se media muestral $\bar{x}$, definidos por el término $(x-\bar{x})^2$. Este término, así como las medias muestrales $\bar{x}$ e $\bar{y}$, se ven notablemente influidos por la representatividad de la muestra respecto de la población, factor condicionado a su vez por el método de selección de la muestra.

En la construcción de modelos se presta notable atención a verificar el cumplimiento de los supuestos básicos antes mencionados, obviando el impacto que el tamaño de la muestra y la representatividad y método de selección de la misma puedan tener en la estimación. La consideración de ambos factores es capital a la hora de diseñar el muestreo para la elaboración de un modelo, pues interesa identificar el tamaño y método de selección de muestra que nos permita obtener una estimación eficiente de los parámetros poblacionales al menor coste posible. En el caso de variables de interés forestal, normalmente costosas de medir, no existen estudios centrados en este tema, aunque como solución de compromiso habitual se sugiere el construir modelos a partir de muestras que abarquen una representación equilibrada de los distintos valores dentro del rango de la variable de interés (muestreo dirigido), y un número de observaciones mínimo superior a 50-60.

El objetivo del presente trabajo es comparar el efecto que los tres tipos de selección muestral más habituales para la construcción de modelos tienen sobre los estimadores de los parámetros y de sus varianzas asociadas en el caso del cumplimiento del resto de supuestos básicos de regresión. Los tipos de selección comparados han sido (1) *aleatoria*; (2) *dirigida*; y (3) *proporcional*. Las comparaciones se han realizados sobre una población bivariante simulada, aplicando métodos de Monte Carlo para la extracción de las muestras, y contrastando el sesgo y precisión en la estimación de los parámetros muestrales respecto de los parámetros poblaciones conocidos.

MATERIAL Y MÉTODOS

Población

Se ha procedido a generar una población de tamaño N = 1.000.000 de observaciones bivariantes (x, y) (figura 1). Para ello en primer lugar se han simulado 1.000.000 de realizaciones de una variable aleatoria x normal de media cero y varianza 100.00. A partir de estas observaciones se ha generado una variable y, relacionada con x por medio de la siguiente función lineal:

$$y = 100 + 0.5\,x + e$$

Donde e es una realización de una variable aleatoria normal de media cero y varianza 1.000. De acuerdo a la expresión anterior, 100 y 0,5 representarían los parámetros poblacionales β_0 y β_1 que caracterizan la relación entre las variables x e y.

Selección de la muestra y ajuste del modelo

A partir de la población simulada se propone usar métodos de Monte Carlo para seleccionar muestras de acuerdo a los siguientes criterios de selección (figura 2):

- *Aleatoria*, donde todas las observaciones poblacionales tienen la misma posibilidad de ser elegidas, imitando la muestra la distribución de frecuencias de la población
- *Dirigida*, las observaciones muestrales se eligen al objeto de tener una representación equilibrada dentro del rango de los valores de la variable explicativa x. Este tipo de selección es muy habitual en la construcción de modelos forestales, donde se busca una representación equilibrada dentro del rango de valores de la variable de interés, sin considerar su representatividad.
- *Proporcional*: la probabilidad de selección es proporcional a la variable explicativa x. Un ejemplo de este tipo de selección muestral viene dado por el uso de datos del Inventario Forestal Nacional español, donde la probabilidad de selección de los árboles en la muestra es proporcional al diámetro de los mismos. En nuestro caso de simulación, si p es la probabilidad de selección aleatoria de una observación cualquiera:
  - Si x <50, la probabilidad de selección es 0.25 p
  - Si 50 < x <150, la probabilidad de selección es p
  - Si 150 < x <250, la probabilidad de selección es 5 p
  - Si x >250, la probabilidad de selección es 50 p

Para cada tipo de selección se generan 2.000 muestras de tamaño n = 700, ajustándose ajustar el modelo de regresión lineal simple por mínimos cuadrados ordinarios a cada una de ellas, y estimándose para cada muestra i el valor medio e intervalos de confianza al 95% para los pará-

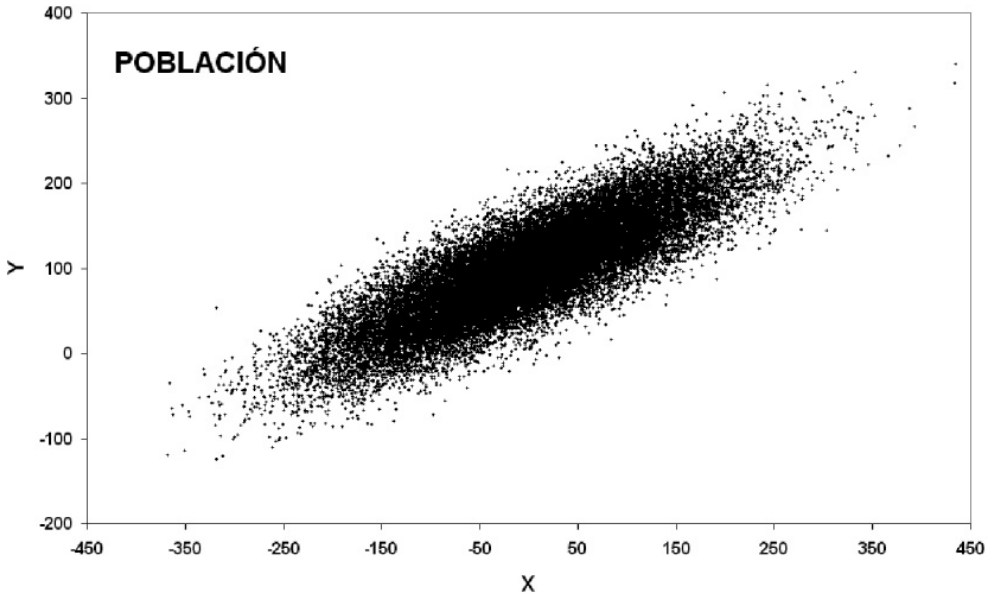

***Figura 1**. Representación bivariante de la población simulada*

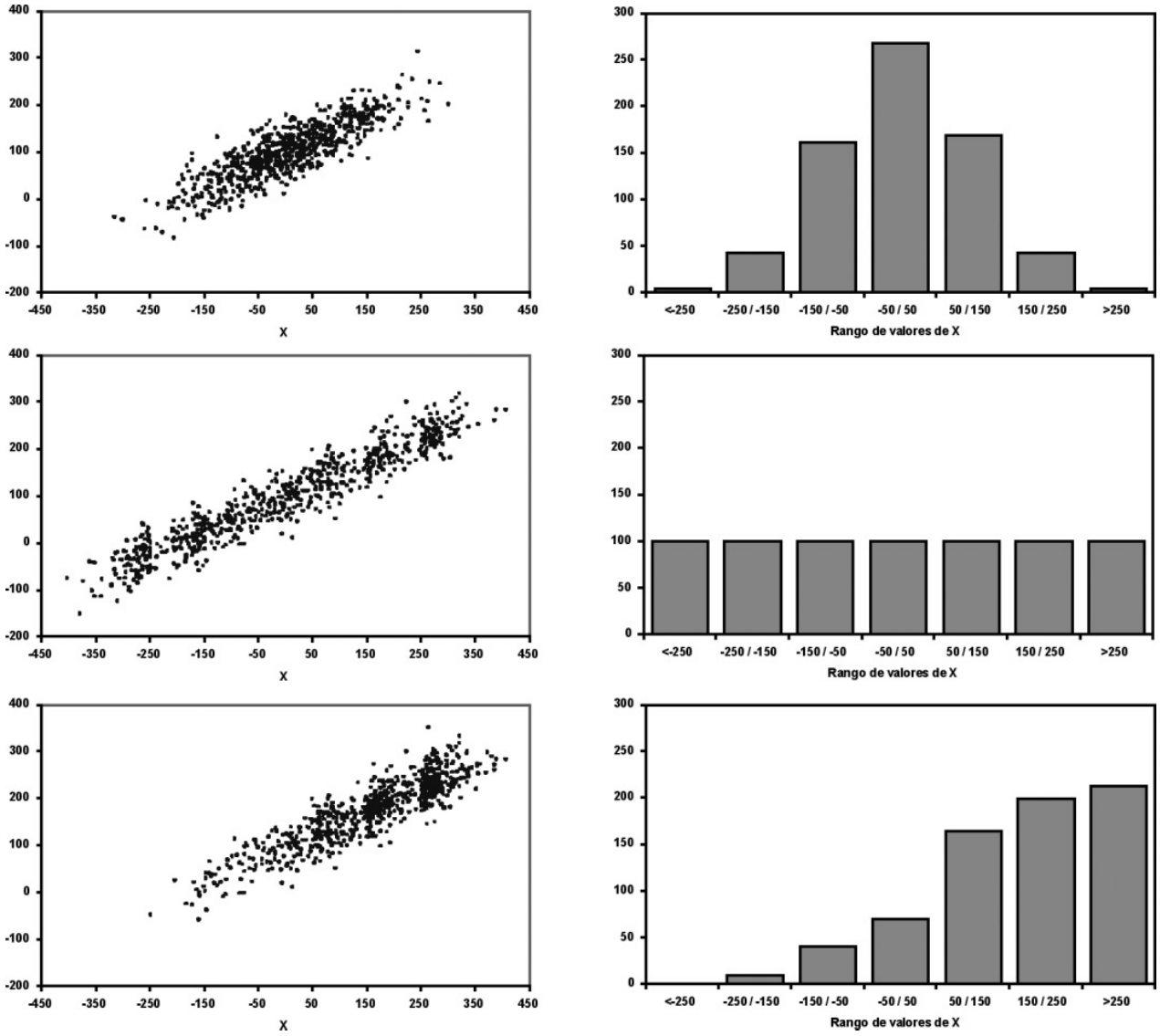

Figura 2. *Representación bivariante (izda) e histograma de frecuencias de valores de x (derecha) de una muestra seleccionada de la población de la Figura 1 de forma aletoria (a), dirigida (b) y proporcional al valor de x (c)*

metros muestrales b_{0i} y b_{1i}, así como el error estándar de los mismos ($\sqrt{S^2(b_{0i})}$ y $\sqrt{S^2(b_{1i})}$).

Comparación de métodos

A partir de los estimadores obtenidos, se plantea comparar los tres métodos de selección muestral a partir de los siguientes estadísticos:

– *Sesgo* de los estimadores:

Asimismo, se determina el *nivel de significación* asociado al valor del *sesgo* mediante un test *t-student.*

$$\text{Sesgo}(b_0) = \sum_{i=1}^{2000} \frac{(b_{0i} - 100)}{2000} \qquad \text{Sesgo}(b_1) = \sum_{i=1}^{2000} \frac{(b_{1i} - 0.50)}{2000}$$

– *Precisión de los estimadores*: valor medio y percentiles 5 – 95 del error estándar de estimación de cada parámetro

– *Significación del estimador*: % realizaciones en las que el intervalo de confianza al 95% para el estimador del parámetro no contiene al valor real del mismo

– *REMC*: valor medio de la raiz del error medio cuadrático

Tanto la generación de la población como las sucesivas selecciones muestrales y el cálculo de los estadísticos de comparación se han realizado a través de macros y rutinas que utilizan procedimientos del módulo STAT de SAS® (ver FAN et *al.*, 2002).

RESULTADOS

La Tabla 1 muestra los resultados de la comparación de los estadísticos antes definidos para los tres tipos de muestreo, sobre 2.000 muestras de tamaño n = 700 seleccionadas mediante métodos de Monte Carlo a partir de la muestra poblacional simulada.

DISCUSIÓN Y CONCLUSIONES

El muestreo aleatorio es el único que garantiza estimaciones insesgadas de ambos parámetros, o lo que es lo mismo, que el promedio de las estimaciones de los parámetros no es significativamente distinto (p-valor >0,05) al valor real de los mismos, La estimación del parámetro b0 se ve menos afectada por un tipo de muestreo dirigido o proporcional que la del parámetro b1, donde estos tipos de selección muestral tienden a sobreestimar de forma significativa el valor del parámetro muestral.

En cuanto a la precisión del estimador, definido a partir del valor medio y distribución de los errores estándares de estimación de los parámetros muestrales, se observa que en el caso de b1 existen diferencias significativas entre los tres tipos de selección, siendo mayores en el caso de la selección muestral aleatoria y menores en la dirigida, Esto se debe a que en este último tipo de muestreo se produce un incremento en el término del numerador de la ecuación [4] $\Sigma(x_i - \bar{x})^2$, debido al aumento de la dispersión de la variable x_i, al incrementarse el número de observaciones en las clases más distanciados respecto a $\bar{x}$, En el caso del parámetro b0, los errores estándar de estimación son mayores en el muestreo proporcional, puesto que este estimador depende de forma directa de $\bar{x}$(ec, [3]), y un aumento de la probabilidad de selección hacia los mayores valores de x produce un incremento del valor medio de la variable explicativa, En el caso de los muestreos dirigido y aleatorio, no se produce variación en el valor de $\bar{x}$.

El porcentaje de ocasiones en los que el intervalo de confianza para los parámetros no incluye al valor real del mismo, es del orden del 4,80 - 5,30% para los dos parámetros y tres tipos de muestreo, excepto en la estimación del parámetro b_1 mediante una selección de tipo aleatorio, donde alcanza un porcentaje del 6%, Debe tenerse en cuenta que el porcentaje de casos esperados al azar sería del 5%, siendo poco significativas las desviaciones respecto a este valor, Por último, los menores valores de REMC corresponden a los tipos de muestreo aleatorio y dirigido.

De acuerdo a los resultados anteriores, y para el caso de poblaciones normales y regresión lineal simple, el método de selección más

			Tipo de muestreo		
	Estadístico		***Aleatorio***	***Dirigido***	***Proporcional***
Parámetro b_0	**Sesgo**		-0,0079	-0,0583	0,1073
	p-valor sesgo		0,7652	0,0311	0,0137
	Error estándar	**Medio**	1,1965	1,1963	1,9159
		p5	1,1447	1,1453	1,8286
		p95	1,2472	1,2473	2,0017
	% No significación		5,30%	5,30%	5,20%
Parámetro b_1	**Sesgo**		0,0005	-0,0007	-0,0014
	p-valor sesgo		0,0928	<0,0001	<0,0001
	Error estándar	**Medio**	0,0120	0,0063	0,0098
		p5	0,0112	0,0060	0,0093
		p95	0,0127	0,0066	0,0103
	% No significación		6,00%	4,85%	4,90%
REMC	**Valor medio**		31,6341	31,6524	31,6927

Tabla 1*. Principales resultados, Donde REMC: raíz del error medio cuadrático, p5 y p95: percentiles 5 y 95 de la distribución de errores estándar*

recomendable es el aleatorio, siempre que el tamaño de muestra garantice la representación de todos los valores existentes en el rango de la variable explicativa *x*, ya que permite obtener estimaciones insesgadas del parámetro poblacional, La selección por muestreo proporcional resulta la menos adecuada, debiendo plantear correcciones que ponderen cada observación por su probabilidad de selección, Por último, la selección dirigida parece una alternativa de compromiso aceptable, especialmente en muestras pequeñas, al garantizar la representatividad de todo el rango, obteniéndose los menores errores estándar de estimación para el parámetro asociado a la variable explicativa.

En el presente trabajo se demuestra la potencialidad del uso de los métodos de Monte Carlo para el estudio de la muestra óptima para modelización, tal y como se ha propuesto en trabajos anteriores (p.ej. FAN et *al.*, 1999). En el caso de la modelización forestal, el presente trabajo puede ampliarse a casos de estudio más habituales, como son la regresión lineal múltiple, regresión no lineal, y regresión con incumplimiento de los supuestos básicos, Asimismo, estas técnicas de simulación de poblaciones y remuestreo deben utilizarse en la identificación del tamaño óptimo de muestra.

Agradecimientos

El presente trabajo se ha desarrollado en el marco financiero y funcional de los proyectos AGL-15521 "*Dinámica de masas heterogéneas de P, pinea: de la respuesta fisiológica a la modelización a escala regional en un escenario de cambio global*" y PSE-31000-2,3 "*Restauración y Gestión Forestal: DECIDE*".

BIBLIOGRAFÍA

FAN, X.; THOMPSON, B. & WANG, L.; 1999. The effects of sample size, estimation methods and model specification on SEM fit indices, *Structural Equation Modeling: a multidisciplinary journal* 6: 56-83

FAN, X.; FELSOVALYI, A.; SIVO, S.A. & KEENAN, S.C.; 2002. *SAS® for Monte Carlo Studies, A guide for quantitative researchers.* SAS Institute Inc. Cary. NC.

MYERS, R.H.; 1990. *Classical and modern regression with applications.* (2nd Edition). Duxbury Classical Series. Pacific Grove. CA.

SEARLE, S.R.; 1971. *Linear Models.* Wiley. New York.