La estimación no paramétrica de curvas con datos sesgados

Noelia Sánchez Martínez

Curso 2024-2025

Por motivos de confidencialidad no es posible la publicación de la memoria completa del Trabajo Fin de Máster titulado "La estimación no paramétrica de curvas con datos sesgados", por lo que en el presente documento se incluye un resumen del mismo.

Resumen

En el ámbito de la Estadística, uno de los supuestos más comunes es que las muestras disponibles son representativas de la población objetivo, lo que permite inferir con fiabilidad propiedades poblacionales a partir de los datos observados. Sin embargo, esta condición no siempre se cumple en la práctica. En numerosos contextos, el procedimiento de muestreo introduce un sesgo sistemático, favoreciendo la inclusión de ciertos individuos y dificultando la de otros, lo que da lugar a datos sesgados. Uno de los casos de más interés en la literatura ocurre cuando la probabilidad de selección depende del valor de la variable que se desea estudiar, como ocurre en el caso del sesgo longitudinal, donde dicha probabilidad es directamente proporcional al valor de la variable.

Este trabajo de fin de máster se centra en el estudio de la estimación no paramétrica de la función de densidad y de la función de distribución a partir de muestras con sesgo longitudinal. A lo largo de los distintos capítulos se presentan ejemplos reales de datos sesgados, se revisan metodologías existentes, se proponen nuevos desarrollos y se implementan estas herramientas en un paquete de *software* específico.

La presencia de un muestreo sesgado impide la aplicación directa de muchos métodos estadísticos clásicos, especialmente aquellos de naturaleza paramétrica, ya que los datos observados no siguen la distribución poblacional objetivo. Por ello, resulta fundamental el desarrollo de metodologías específicas que tengan en cuenta

la función de sesgo, y en particular, de técnicas no paramétricas que permitan estimar de forma robusta las funciones poblacionales de interés. Esta necesidad, unida al interés práctico por disponer de herramientas aplicables en contextos reales, justifica el enfoque metodológico y computacional adoptado en este trabajo.

El Capítulo 1 introduce el concepto de datos sesgados, ilustrándolo con ejemplos procedentes de diferentes ámbitos como la Biología, la Medicina, la Industria o la Economía. Uno de los ejemplos clásicos es el presentado por Cox (2005), relacionado con la longitud de fibras textiles. En ese caso, las fibras más largas tienen mayor probabilidad de ser seleccionadas, lo que genera una muestra sesgada. Este tipo de sesgo, conocido como sesgo longitudinal, ha sido ampliamente estudiado en la literatura, tanto en lo relativo a la estimación de la función de densidad (Bhattacharyya et al. (1988); Jones (1991); Borrajo et al. (2017)) como a la estimación de la función de distribución (Cox (2005); Bose y Dutta (2022)).

Además, en este capítulo se describe con detalle un conjunto de datos reales que es utilizado a lo largo del trabajo como hilo conductor: se trata de mediciones del ancho de arbustos de la especie *Cercocarpus montanus*, recogidas mediante métodos de transecto lineal en una antigua cantera de piedra caliza en Wyoming (EE.UU.). Dado el procedimiento de muestreo empleado, los arbustos más anchos tenían mayor probabilidad de ser seleccionados, lo que da lugar a un claro ejemplo de muestra con sesgo longitudinal. Este conjunto de datos, recogido en el otoño de 1986 por estudiantes de un curso de posgrado impartido por Lyman L. McDonald, constituye una base empírica valiosa para ilustrar los métodos estudiados. Pueden consultarse más detalles del muestreo, así como las mediciones para el resto de características en Muttlak (1988).

El Capítulo 2 está dedicado a la estimación no paramétrica de la función de densidad a partir de datos con sesgo longitudinal. Es importante notar que en este contexto, no se busca estimar la función de densidad que genera los datos observados, sino la función de densidad que generaría los valores insesgados que no se observan, lo que deriva en un incremento de la complejidad de los métodos a emplear. A lo largo del capítulo se revisan los principales estimadores propuestos en la literatura, como el estimador de Bhattacharyya et al. (1988) o el estimador tipo núcleo introducido por Jones (1991). De manera análoga a lo que ocurre en datos insesgados, uno de los retos de los métodos tipo núcleo es la elección del parámetro de suavizado. Para ahondar en esta cuestión, se estudian distintos selectores de este parámetro, incluyendo la propuesta de Guillamón et al. (1998) basada en validación cruzada sobre el error cuadrático integrado (ISE, *Integrated Squared Error*), una regla del pulgar y dos selectores *bootstrap* desarrollados por Borrajo et al. (2017) que se apoyan en la parte dominante del error cuadrático medio integrado (MISE, *Mean Integrated Squared Error*). Todos estos métodos son aplicados al conjunto de datos sobre *Cercocarpus montanus* para ilustrar sus características prácticas y sus diferencias de comportamiento.

El Capítulo 3 se centra en la estimación no paramétrica de la función de distribución, prestando atención tanto a resultados teóricos como a desarrollos metodológicos innovadores. En primer lugar, se deriva una expresión analítica del MISE para el estimador tipo núcleo de la función de distribución propuesto por Bose y Dutta (2022), y sobre esta base se construyen tres nuevos selectores globales del parámetro de suavizado: una regla del pulgar, una extensión del criterio de validación cruzada de Bowman et al. (1998), y un selector *plug-in* basado en la correcta estimación de la integral del cuadrado de la derivada segunda de la función de distribución, para lo que se ha necesitado determinar la expresión de su error cuadrático medio (MSE, *Mean Squared Error*).

A diferencia del selector local ya existente en este contexto que fue presentado por Bose y Dutta (2022), las nuevas propuestas de selección del parámetro de suavizado no requieren la elección de constantes adicionales. El capítulo incluye además un extenso estudio de simulación que compara el rendimiento de los distintos selectores propuestos, mostrando que los nuevos métodos son competitivos con el selector local existente. Al igual que en el Capítulo 2, los resultados se ilustran mediante la aplicación de los métodos anteriormente descritos a los datos de arbustos.

La implementación computacional de métodos estadísticos representa un paso crucial para su transferencia desde el ámbito teórico hacia la aplicación práctica. La disponibilidad de herramientas de *software* accesibles y bien documentadas no solo facilita el uso de estos procedimientos por parte de la comunidad científica, sino que también acelera su adopción y eventual refinamiento a través de la aplicación en casos reales.

En el contexto específico de la estimación no paramétrica, tanto de la función de densidad como de la función de distribución, existen múltiples paquetes en el entorno de programación R que permiten trabajar con datos insesgados, que han demostrado ser herramientas robustas y eficientes para el análisis de datos. Sin embargo, cuando los datos presentan algún tipo de sesgo en el proceso de muestreo, los métodos tradicionales quedan invalidados y, por ende, estos paquetes ya no resultan de utilidad para el tratamiento de los mismos.

A pesar de lo habituales que son los datos sesgados en la práctica y de los desarrollos teóricos existentes en la literatura, no existía hasta la fecha ningún paquete de que implementara de manera integral los métodos de estimación no paramétrica específicamente diseñados para datos sesgados. El Capítulo 4 aborda el desarrollo y presentación del paquete `WData` (*Weighted Data*), una herramienta en lenguaje R diseñada para implementar de forma integrada los métodos de estimación estudiados. `WData` permite aplicar de forma sencilla los estimadores y selectores presentados en los capítulos anteriores y, además, incluye una función para generar muestras sesgadas, útil para estudios de simulación.

Por otra parte, el sesgo longitudinal es muy habitual en la práctica, pero existe una gran variedad de marcos de muestreo que incurren en otras funciones de sesgo. Si

bien los resultados presentados en este trabajo hacen referencia al sesgo longitudinal, algunos de los estimadores han sido extendidos en la literatura a otras funciones de sesgo. Con relación a la función de densidad, la extensión del estimador tipo núcleo puede consultarse en Jones (1991) y la expresión de su MSE y MISE en Borrajo et al. (2017). En lo que se refiere a la función de distribución, el estimador empírico de Cox (2005) fue extendido por Efromovich (2004). Aunque este trabajo se centra en el caso del sesgo longitudinal (función de sesgo $w(x) = x$), el paquete ha sido diseñado para aceptar cualquier función de sesgo que sea positiva, acotada e integrable sobre el soporte de los datos. Esto lo convierte en una herramienta flexible, apta para ser extendida a nuevos contextos y problemas estadísticos.

El Capítulo 5 recoge las principales conclusiones del trabajo y plantea posibles líneas futuras de investigación. Entre ellas se incluye la extensión de los métodos desarrollados a la estimación de otras funciones poblacionales como la función cuantil o la función de regresión, así como la adaptación de los procedimientos al contexto multivariante. Por ejemplo, el estimador empírico de Cox (2005) podría servir como base para definir una función cuantil empírica en presencia de sesgo, y el trabajo de Ahmad (1995) constituye una referencia relevante en el estudio del caso multivariante. En definitiva, aunque este trabajo se ha centrado en problemas univariantes y en el sesgo longitudinal, muchas de las ideas desarrolladas tienen potencial para aplicarse en contextos más generales.

Bibliografía

Ahmad, I. A. (1995). On multivariate kernel estimation for samples from weighted distributions. *Statistics & Probability Letters*, 22, 121-129.

Bhattacharyya, B. B., Franklin, L. A., y Richardson, G. D. (1988). A comparison of nonparametric unweighted and length-biased density estimation of fibres. *Communications in Statistics - Theory and Methods*, 17, 3629-3644.

Borrajo, M. I., González-Manteiga, W., y Martínez-Miranda, M. D. (2017). Bandwidth selection for kernel density estimation with length-biased data. *Journal of Nonparametric Statistics*, 29, 636-668.

Bose, A., y Dutta, S. (2022). Kernel based estimation of the distribution function for length biased data. *Metrika*, 85, 269-287.

Bowman, A., Hall, P., y Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85, 799-808.

Cox, D. (2005). Some sampling problems in technology. En D. Hand y A. Herzberg (Eds.), *Selected Statistical Papers of Sir David Cox* (pp. 81-92, Vol. 1). Cambridge University Press.

Efromovich, S. (2004). Distribution estimation for biased data. *Journal of Statistical Planning and Inference*, 124, 1-43.

Guillamón, A., Navarro, J., y Ruiz, J. M. (1998). Kernel density estimation using weighted data. *Communications in Statistics - Theory and Methods*, 27, 2123-2135.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika*, 78, 511-519.

Muttlak, H. A. (1988). *Some aspects of ranked set sampling with size biased probability of selection* [Tesis doctoral, University of Wyoming].