

Data Analysis Using Regression and Multilevel/Hierarchical Models.

Andrew Gelman y Jennifer Hill (2007). Cambridge, UK: Cambridge University Press.

*Luis Maldonado**

En cursos de métodos estadísticos avanzados, habitualmente son utilizados textos econométricos como principal referencia. Sin embargo, econometría es sólo una perspectiva a partir de la cual se puede aprender estadística. Una alternativa la ofrece la disciplina de la estadística misma, en donde conceptos tales como variabilidad de los coeficientes, incerteza en torno a dicha variabilidad o simulación, son de uso común en la actualidad. Lamentablemente, la discusión en torno a dichos conceptos es frecuentemente planteada en términos altamente matemáticos, lo cual hace difícil el estudio de métodos avanzados y los conceptos usados por los estadísticos para el promedio de las personas formadas en las ciencias sociales.

En este contexto, el libro *Data Analysis Using Regression and Multilevel/Hierarchical Models*, publicado por Andrew Gelman y Jennifer Hill en 2007, merece un lugar especial en la biblioteca de investigadores interesados en la aplicación de métodos estadísticos avanzados en ciencias sociales. Gelman y Hill han hecho un gran trabajo para presentar y discutir de un modo intuitivo técnicas estadísticas avanzadas, en particular la técnica cuantitativa, que es la que con mayor fuerza se ha posicionado en la actualidad en las ciencias sociales: análisis de regresión multinivel/jerárquico, es decir, modelos que incluyen información de unidades que tienen diferentes niveles de agregación (por ejemplo, estudiantes y colegios). Específicamente, el objetivo de los autores es ajustar, graficar y entender el modelo lineal estándar (OLS/MICO) y regresiones multinivel, así como modelos lineales generalizados. Un rasgo distintivo del libro, que puede ser de gran utilidad para el público de las ciencias sociales, es la introducción en el problema del modelamiento estadístico a través de ejemplos, la gran mayoría extraídos de las ciencias políticas y la salud pública. Para entender los contenidos del texto se requiere estadística básica y un conocimiento de regresión lineal. He usado el texto de Gelman y Hill en cursos de posgrado como apoyo a otros textos más básicos y ha sido muy estimulante para alumnos interesados en aprender las técnicas mencionadas en esta reseña. Otros rasgos distintivos del texto son sugerir gráficos como una forma

* Doctor en economía y ciencias sociales, Universidad de Colonia, Alemania. Profesor asistente, Instituto de Sociología, Pontificia Universidad Católica de Chile. lmaldona@uc.cl.

de presentar los resultados de análisis de regresión y enfatizar el uso del software a través de la ilustración de los comandos, particularmente de R y BUGS, lo cual es de gran ayuda para las personas interesadas en estimar los ejemplos del libro en sus computadores personales.

El libro se compone de tres partes más tres apéndices. La primera parte se divide en dos secciones. En la primera sección, Gelman y Hill explican la estimación de modelos de regresión lineal (OLS/MICO) y de regresión logística. El último capítulo de esta sección está dedicado a modelos lineales generalizados, el cual entrega los conceptos claves en los cuales se basa el análisis multinivel. La segunda sección está dedicada a mostrar el uso de modelos de regresión para estimar cantidades de interés sustantivo, en particular predicciones sobre la base de simulaciones. Adicionalmente se discute la estimación de efectos causales en los dos últimos capítulos de esta sección, en donde se muestra cómo una regresión lineal puede ser usada para la estimación de efectos causales.

La segunda parte se focaliza en la estimación e interpretación de modelos de regresión multinivel. Esta parte también contiene dos secciones. La primera sección comienza con la discusión de bases de datos con estructuras de dos niveles, poniendo el foco en la combinación (*pool*) de información referida a la varianza entre las unidades de nivel más agregado e información relacionada con la varianza al interior de cada una de estas unidades. Luego se desarrolla el modelo de regresión lineal multinivel desde lo más simple, es decir, modelos con interceptos aleatorios, para avanzar hacia modelos más complejos representados por modelos con coeficientes aleatorios y *non-nested models*. Los dos últimos capítulos de esta sección discuten modelos lineales generalizados multinivel (logístico binario, ordinal y modelos Poisson). La segunda sección de la segunda parte discute la estimación de modelos multinivel a través del uso de métodos bayesianos, lo cual es ilustrado a través de ejemplos en R y BUGS. Esta sección puede ser una excelente oportunidad para introducirse en la estadística bayesiana, especialmente interesante para personas que dudan de que la realidad social debe ser estudiada a partir de criterios objetivos, independientes de la subjetividad del investigador.

La tercera parte del libro de Gelman y Hill está dedicada a temas que si bien no son esenciales para el modelamiento estadístico, de seguro son de gran utilidad para el investigador aplicado. Dichos temas son la estimación del tamaño muestral, comparación de modelos y de sus ajustes, análisis de varianza (ANOVA), inferencia causal usando modelos multinivel y métodos de imputación de datos perdidos. Finalmente, al término del libro los autores presentan tres apéndices, dentro de los cuales destaca el tercero. Este presenta el código para estimar los modelos discutidos en el texto en otros softwares (por ejemplo, Stata o SAS).

Si bien Gelman y Hill han escrito un excelente libro sobre análisis de regresión, el texto podría ser mejorado en algunos aspectos. Desde un punto de vista técnico, sería recomendable que una futura edición ilustrara la construcción de gráficos en paquetes de R, tales como lattice o ggplot2. Además, como sociólogo me hubiera gustado encontrar una discusión más en profundidad de modelos teóricos de las ciencias sociales y sobre cómo dichos modelos pueden ser evaluados por parte de las técnicas discutidas en el libro, especialmente en qué medida el análisis multinivel sirve para estudiar uno de los temas que más concentran actualmente la atención en las ciencias sociales: la heterogeneidad causal asociada a la contextualización de la realidad social. En suma, Gelman y Hill han escrito un excelente libro que ofrece una gran oportunidad para introducirse en los temas cuantitativos que se discuten en la actualidad en las ciencias sociales.