Nota crítica de libro

NOTA CRÍTICA DE LIBRO

Quantitative Social Science. An Introduction

Kosuke Imai
Princeton y Oxford, Princeton University Press
2017
432 páginas

Ángela M. Diblasi

Profesora
Facultad de Ciencias Económicas
Universidad Nacional de Cuyo

Cuando en 1977 John Tukey afirmó «The greatest value of a picture is when it forces us to notice what we never expected to see» en ocasión de la publicación de su famoso libro *Exploratory Data Analysis,* estaba resumiendo en una expresión el objetivo del Análisis Exploratorio de Datos. También en ese momento, hace ya 40 años, Tukey vislumbraba el potencial que tendría el uso de herramientas computacionales para la manipulación de datos u observaciones de la vida real.

El libro de Kosuke Imai muestra instrumentos del análisis exploratorio de datos aplicados especialmente a las Ciencias Sociales, como la Economía, la Política y la Sociología. Su herramienta computacional es el conocido entorno y lenguaje de programación libre R, el más difundido y actualizado en el mundo de la Estadística.

El texto puede dividirse en tres partes, en función de las temáticas abordadas: el capítulo I, que está dedicado a introducir los conceptos fundamentales para el uso de este software libre; los capítulos II, III, IV y V, que están dedicados específicamente a herramientas del análisis exploratorio, y los capítulos VI y VII, que introducen los conceptos básicos de probabilidad e inferencia estadística, respectivamente.

En el transcurrir de las páginas de este libro se utilizan bases de datos que pueden resultar muy atractivas para las personas interesadas en las Ciencias Sociales y que el autor deja a disposición de sus lectores en su página web (http://press.princeton.edu/qss/). Adicionalmente, cada capítulo presenta una serie de ejercicios muy motivadoa res para profundizar en los conceptos y funciones de R vertidos en él.

En el capítulo I, dedicado a la introducción al lenguaje R, se incorporan operaciones con números, vectores y matrices. También se dedican algunas páginas a la presentación de comandos relativos a resumir archivos mediante tablas y estadísticos básicos y se define el concepto de *objeto* en R. Termina con una cantidad importante de ejercicios relativos a la utilización de los comandos mostrados en el capítulo.

El capítulo II se centra en el estudio del concepto de *causalidad*. Se analiza, mediante ejemplos, el riesgo de inferir causalidad sin realizar un análisis adecuado. Se plantean ejemplos de diseños que posibilitan deducir que un fenómeno es causado por otro.

En el capítulo III se analiza la dificultad de realizar mediciones en el contexto de las Ciencias Sociales. En particular, se examinan los datos de una encuesta sobre la percepción del daño ocasionado por los talibanes y el ejército norteamericano a la población afgana. Se muestran funciones de R para tratar con datos perdidos. Se estudia el sesgo en las estimaciones por efecto de la no respuesta individual a la encuesta, la no respuesta a una pregunta de la encuesta o la falta de veracidad en las respuestas. Se plantea también la realización de algunos gráficos (con sus correspondientes códigos R) para investigar la distribución, tales como histogramas, gráficos de barras y gráficos cuantil-cuantil. Se introducen también las definiciones de índice de Gini y de correlación entre dos variables. Finalmente, se realiza una introducción a la idea de agrupamientos (clustering) y se ilustra con la metodología de las k-medias.

El capítulo IV está dedicado a la predicción. Se introduce el concepto de *predicción* en Estadística mediante el estudio de un ejemplo vinculado a los resultados de las elecciones presidenciales en EEUU, en las que resultó triunfador Obama. Con la motivación de predecir resultados en la elección de 2008 utilizando la del 2004, se introducen los comandos condicionales y los que se utilizan para realizar *loops* en R. Se incorpora también el concepto de regresión, mediante el mismo ejemplo, y se desarrolla el álgebra del método de mínimos cuadrados para estimar los coeficientes del

modelo lineal. Se analizan modelos lineales con factores y sus interacciones.

El capítulo V está dedicado a tres grandes tópicos: el análisis de textos, de redes y de datos espaciales. Para el análisis de textos se utiliza una base de datos llamada The Federalist, colección de artículos y ensayos escritos por Alexander Hamilton, James Madison y John Jay, quienes promueven la ratificación de la Constitución de Estados Unidos. El problema que se plantea es la distinción de los artículos escritos por cada uno de estos autores. Se sabe que algunos de ellos pertenecen a uno u otro, pero no está clara la autoría en otros casos, aunque se conoce que fueron escritos por algunas de las personas citadas. Con el objetivo de predecir dicha autoría, se consideran una serie de funciones de R preparadas en paquetes específicos que permiten contar palabras utilizadas en distintos textos de un corpus, comparar frecuencias y, en consecuencia, predecir autorías. Una de las metodologías adicionales para la validación de estas predicciones es la de validación cruzada *(cross validation*).

En este capítulo se analizan también datos donde lo que interesa es la relación entre ellos más que el valor individual (grafos). Se analizan dos ejemplos, uno relativo a los matrimonios entre las familias poderosas de la Florencia medieval y otro que contiene datos de la red social *Twitter*, referidos a los seguidores de los senadores del congreso de EEUU. Finalmente, se introducen datos espaciales y espacio-temporales de distinta naturaleza y se muestran ejemplos de funciones de R para ubicarlos en mapas.

En el capítulo VI se introducen los conceptos clásico y bayesiano de la idea de probabilidad, el álgebra de la probabilidad, combinatoria, probabilidad condicional e independencia. Se analiza un ejemplo de predicción de la raza de un habitante de EEUU utilizando su nombre y lugar de residencia mediante el estudio de un archivo de datos con esta información y mediante el uso de probabilidades condicionadas con adecuados comandos de R. Se introducen también los conceptos de variable aleatoria y distribución de probabilidad. Se analizan algunas distribuciones discretas y continuas particulares y sus respectivos comandos. Además, se presentan los conceptos de esperanza, varianza, ley de los grandes números y teorema central del límite. Se ilustran algunos conceptos mediante el uso de funciones de R para simulación.

En este mismo capítulo se analizan también los conceptos de estimación y algunas propiedades de los estimadores, como el sesgo y el error estándar. También se plantean intervalos de confianza y se evalúa la determinación de un adecuado tamaño muestral. Se introducen, asimismo, el concepto de hipótesis estadística y conceptos relativos, como tamaños de errores, el p-valor y la potencia de un test.

Otra noción de este capítulo es la referida a los modelos de regresión con variables (explicativas) aleatorias. Se dedica una parte a la estimación de coeficientes, sus propiedades, supuestos y análisis de los supuestos.

Finalmente, quiero destacar lo que es, desde mi punto de vista, el aporte más interesante de este libro: la puesta a disposición del lector de una muy abundante cantidad de bases de datos provenientes de investigaciones realizadas con observaciones de la vida real en el campo de las Ciencias Sociales y las correspondientes herramientas de R para analizarlas. Recomiendo su lectura a los estudiantes de Ciencias Sociales e incluso a investigadores que quieran introducirse en ese invaluable, exhaustivo y actualizado conjunto de instrumentos exploratorios de datos que ofrecen las miles de herramientas desarrolladas y puestas a disposición por el equipo de R.