

Una introducción a R para la investigación en Ciencias Sociales

José Julián Escario y Juan Valiño

PRENSAS DE LA UNIVERSIDAD DE ZARAGOZA

*UNA INTRODUCCIÓN A R
PARA LA INVESTIGACIÓN EN CIENCIAS SOCIALES*

José Julián Escario y Juan Valiño

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Dirijase a CEDRO (Centro Español de Derechos Reprográficos, www.cedro.org) si necesita fotocopiar o escanear algún fragmento de esta obra.

© José Julián Escario y Juan Valiño
© De la presente edición, Prensas de la Universidad de Zaragoza
(Vicerrectorado de Cultura y Proyección Social)
1.^a edición, 2020

Colección de Textos Docentes, n.º 296

Prensas de la Universidad de Zaragoza. Edificio de Ciencias Geológicas, c/ Pedro Cerbuna, 12, 50009 Zaragoza, España. Tel.: 976 761 330. Fax: 976 761 063
puz@unizar.es <http://puz.unizar.es>



Esta editorial es miembro de la UNE, lo que garantiza la difusión y comercialización de sus publicaciones a nivel nacional e internacional.

ISBN 978-84-1340-158-4

Impreso en España

Imprime: Servicio de Publicaciones. Universidad de Zaragoza

D.L.: Z 1033-2020

Prólogo

A la hora de realizar tanto trabajos de fin de grado (TFG) como artículos de investigación en el campo de las ciencias médicas y sociales, el análisis de datos cada vez adquiere más relevancia debido a la creciente disponibilidad de bases de datos proporcionadas por muchos organismos oficiales. Además, también los estudiantes e investigadores diseñan frecuentemente sus propias encuestas o experimentos para obtener los datos relevantes para sus estudios.

Para analizar los datos es necesario disponer de herramientas estadísticas adecuadas. Un problema relacionado con dichas herramientas (SPSS, Stata, SAS, etc.) es que suelen ser bastante costosas. Por lo tanto, los alumnos matriculados en la asignatura Trabajo Fin de Grado rara vez pueden permitirse su compra. El inconveniente anterior puede llevar a los alumnos a tratar de conseguir copias «piratas» de dichos programas. A veces, incluso los profesores sugieren esa forma de conseguir los programas. Sin embargo, esta práctica plantea dos problemas. El primero es de tipo ético y legal, puesto que supone una adquisición ilegal que vulnera la ley de propiedad intelectual. El segundo consiste en que puede afectar a la seguridad informática del usuario. Así, los programas pueden contener un código malicioso desarrollado por las personas que los han desprotegido.

Por todo ello, sería interesante que tanto los profesores como los alumnos que desarrollan trabajos de fin de grado o estudios de investigación pudieran disponer de una herramienta estadística gratuita. Una opción interesante es el programa estadístico R. Este programa, cuyo uso se ha expandido con mucha fuerza en los últimos años, presenta varias ventajas:

- Permite realizar prácticamente cualquier análisis estadístico.
- Es gratuito.
- Es muy flexible.
- Funciona en diversos sistemas operativos (Windows, Mac, Linux, etc.).
- Se han desarrollado entornos gráficos (R Commander, Rattle, RStudio, etc.) que facilitan su manejo y que lo convierten en una herramienta sumamente interesante.

En realidad, R es un lenguaje de programación y un entorno para el análisis de datos que se distribuye bajo la General Public Licence (GPL). Se trata de una implementación moderna del lenguaje de programación S desarrollado por la empresa Bell Laboratories. R fue inicialmente desarrollado en 1993 en la Universidad de Auckland (Nueva Zelanda) por los profesores Robert Gentleman y Ross Ihaka. En la actualidad, es desarrollado por multitud de personas de todo el mundo coordinadas por un equipo de desarrollo (denominado R Core Team Development).

Para nuestros objetivos, podemos considerar R como un programa, ya que funciona igual que cualquier otro programa informático. Así pues, como cualquier programa necesitamos descargarlo e instalarlo en nuestro ordenador para poder utilizarlo. Aunque una vez instalado R, ya se puede realizar una gran cantidad de análisis estadísticos básicos, existe una multitud de paquetes que expanden las posibilidades de R con muchas funcionalidades. Dichos paquetes se encuentran disponibles en una red denominada Comprehensive R Archive Network (CRAN).

Esencialmente, R funciona como un lenguaje de programación. Así, el investigador escribe una serie de comandos, los selecciona y los envía para su ejecución, tras lo cual el programa produce los resultados que se derivan de dichos comandos. Aunque hay entornos o interfaces gráficos que permiten trabajar únicamente mediante ventanas (p. ej. R Commander y Rattle), en este manual trabajaremos con comandos. No obstante, nos ayudaremos de un entorno gráfico, denominado RStudio, que facilita enormemente la escritura de las secuencias de comandos.

El presente manual se ha creado con el objetivo de ayudar a todos aquellos profesores y alumnos que deseen realizar trabajos de investigación y/o trabajos de fin de grado utilizando datos de corte transversal (encuestas, estudios epidemiológicos, etc.). Uno de los autores de este manual, tras usar diversos programas estadísticos (TSP, Stata, SPSS, etc.), lleva los tres últimos años trabajando exclusivamente con R. Su utilización, tanto en artículos de investigación como en la asignatura Trabajo Fin de Grado, le ha permitido constatar que se trata de una herramienta útil en ambos casos. Además, con la adecuada orientación, los alumnos pueden manejar el programa sin problemas. La experiencia así obtenida ha permitido desarrollar el presente documento como un manual de introducción en el que se presenta un conjunto sencillo de instrucciones o comandos para realizar los análisis que se requieren en la mayoría de los trabajos de investigación (carga y lectura de los datos, recodificación de variables, análisis descriptivos, tablas de contingencia, contrastes de independencia de variables, contrastes de igualdad entre medias, estimaciones, tablas con los modelos estimados, etc.). En este manual también se indica cómo se puede presentar en un procesador de textos, bien directamente o mediante

un simple proceso de copiar y pegar, la mayoría de los resultados obtenidos en forma de tablas y gráficos.

En relación con los modelos de regresión, en el presente manual se prestará especial atención al modelo de regresión logística (o modelo logit), al modelo probit y a seis modelos de regresión específicos para datos de conteo. Estos modelos de regresión son ampliamente utilizados en trabajos de investigación tanto en ciencias sociales (economía, sociología, *marketing*, etc.) como en otras ciencias (medicina, biología, ecología, etc.). A este respecto, el manual es una útil guía para:

- Aprender a estimar los modelos considerados.
- Tener una visión intuitiva de cada modelo de regresión considerado.
- Interpretar las estimaciones obtenidas en cada modelo.
- Aplicar varias herramientas para validar y comparar los modelos estimados.
- Valorar el grado de ajuste de las predicciones a los datos observados.

En definitiva, este manual está especialmente dirigido a aquellas personas que quieran introducirse en el uso de R (utilizando RStudio) de una forma rápida y sencilla para poder realizar trabajos de investigación con datos de corte transversal. Se asume que el lector tiene conocimientos básicos de estadística descriptiva y de cómo se interpretan los coeficientes estimados en un modelo de regresión lineal.

Varias características, considerándolas en su conjunto, hacen que este manual venga a llenar un hueco importante. Por un lado, está escrito en castellano. Por otro lado, la estructura es novedosa. Así, frente a otros manuales, no pone el foco en aspectos relacionados con las distribuciones de probabilidad, no pone el foco en la realización de gráficos, no se centra en explicar multitud de modelos de regresión muy diferentes, etc.

El foco está en explicar cómo estimar en R los modelos de regresión más ampliamente utilizados y en cómo interpretar los resultados obtenidos. Obviamente, intentar escribir un manual introductorio breve y conciso, implica que queden fuera muchos temas en los que el lector podría estar interesado, por ejemplo, otros modelos de regresión. No obstante, cuando se empieza en cualquier ámbito de la vida es preferible no intentar abarcar demasiado. Para facilitar el aprendizaje de los contenidos del manual, se proporciona al lector la base de datos utilizada y un fichero con todos los comandos utilizados.¹ De esta forma, el lector podrá reproducir todos

¹ El fichero de comandos y los datos pueden descargarse desde el siguiente enlace: <https://drive.google.com/drive/folders/1o4m7BJ5oQZNafys-noTQDzlwYgYYgTEE?usp=sharing>.

los análisis llevados a cabo, así como experimentar con nuevas variantes y opciones.

Consideramos que el presente manual puede ser de utilidad para todos aquellos que quieran trabajar con encuestas o bases de datos en multitud de ciencias (economía, *marketing*, sociología, psicología, medicina, etc.) y en especial para:

- Los alumnos que realicen trabajos de fin de grado o master analizando datos.
- Los profesores que dirijan dichos trabajos.
- Los alumnos de doctorado y profesores interesados en realizar artículos de investigación con encuestas o bases de datos similares.

Para finalizar con el prólogo, nos gustaría indicar que R va mucho más allá de ser un programa o un lenguaje de programación. Se trata de un gran proyecto colaborativo en el que muchas personas contribuyen desinteresadamente para que todos, ya sea en el ámbito investigador, universitario, empresarial, etc., podamos utilizar una potente herramienta estadística. Dicho de otra forma, se trata de software libre que nos beneficia a todos. Esta idea se refleja perfectamente en una cita de Ross Itaka, uno de los cofundadores de R:²

R cambió mi opinión sobre la humanidad hasta cierto punto, al ver cómo las personas están realmente dispuestas a dar libremente de sí mismas y producir algo más grande que ellas sin pensar en la gloria personal. Hay mucho trabajo sin reconocimiento.

² La frase original es: «R changed my opinion of humanity to some extent, to see how people are really willing to freely give of themselves and produce something larger than themselves without any thought of personal glory. There's a lot of work with no recognition».

ÍNDICE

PRÓLOGO	7
1. INSTALACIÓN DE R Y RSTUDIO	11
1.1. Descarga e instalación.....	11
1.2. Apariencia y uso de RStudio.....	15
2. CARACTERÍSTICAS DE R Y CONCEPTOS BÁSICOS	23
3. LECTURA DE DATOS Y RECODIFICACIÓN DE VARIABLES	33
3.1. Carga y lectura de datos	33
3.2. Recodificación y generación de variables.....	38
4. ANÁLISIS DESCRIPTIVOS	45
4.1. Medidas de posición y dispersión	45
4.2. Tablas de contingencia y gráficos asociados	47
4.3. Contrastes de hipótesis sobre medias y varianzas.....	56
5. ANÁLISIS DE REGRESIÓN CON VARIABLES DEPENDIENTES DICOTÓMICAS.....	61
5.1. Introducción	61
5.2. Estimación	62
5.3. Interpretación de los coeficientes.....	68
5.4. Medidas de la calidad del ajuste de las estimaciones.....	73
6. ANÁLISIS DE REGRESIÓN CON VARIABLES DEPENDIENTES DE CONTEO	81
6.1. Introducción	81
6.2. Estimación	84
6.3. Interpretación de los coeficientes.....	93
6.3.1. Modelos de regresión de Poisson y Binomial negativa II	93
6.3.2. Modelos valla	96
6.3.3. Modelos con inflado de ceros	97
6.4. Medidas de la calidad del ajuste y selección de modelos	98
Abreviaturas	110
Paquetes utilizados	110

ISBN 978-84-1340-158-4



9 788413 401584

SOCIALES



colección
textos docentes



Prensas de la Universidad
Universidad Zaragoza