Instituto Mora
Doctorado en Estudios del Desarrollo

Métodos de Investigación Avanzados
Opción Métodos Cuantitativos

Martín Paladino*

Presentación

El curso se propone como una continuación de la formación en Métodos Cuantitativos que reciben los estudiantes de los programas de posgrado de Instituto Mora. Aprovechando aportes recientes en el campo nos centraremos en métodos aplicables a variables categóricas, nominales u ordinales, de uso frecuente en las ciencias sociales. Los métodos que aplicaremos atienden a dos problemas frecuentes en la investigación social: el de descripción (incluyendo sumarios y gráficos) de datos cuantitativos y el de modelado de las relaciones entre conjuntos de variables heterogéneas (modelos lineales generalizados).

La parte práctica del curso tendrá lugar a partir de la implementación de las técnicas en la plataforma de análisis estadístico R. Adicionalmente cada unidad estará acompañada de la lectura de al menos una investigación reciente en el ámbito de la sociología en la que el método en cuestión sea aplicado. De este modo podremos dar cuenta del proceso de inferencia sustantiva a través del cuál se ponen en relación los resultados del análisis con las hipótesis teóricas de algún campo temático.

La comunicación durante el curso, incluyendo la publicación de los ejercicios semanales, se hará a través del sitio https://metodoscuantitativos2018.netlify.com/

Las guías de R están disponibles en https://martinpaladino.github.io/rsociales/

*Instituto Mora, mpaladino@mora.edu.mx

Objetivos

Al concluir el curso los alumnos podrán:

- Utilizar la plataforma estadística de fuente abierta R para la carga, manejo y análisis de datos cuantitativos.
- Explorar y describir datos categóricos y continuos a través de estadísticos y gráficos adecuados y elegantes.
- Modelar variables dependientes continuas y categóricas (dicotómicas, politómicas y ordinales).

Evaluación del curso

Para la evaluación del curso se tomarán en consideración los siguientes items.

- Asistencia y participación en clase. (20 %)
- Trabajos prácticos. Después de cada sesión se presentará a los alumnos un problema que deberán resolver. (40 %)
- Trabajo final. Los alumnos entregarán como trabajo final un reporte técnico aplicando los métodos que han aprendido. Dicho trabajo estará aplicado a sus proyectos de investigación, ya sea atendiendo de manera directa a sus hipótesis o al menos contextualizando el problema de investigación que están tratando. (40 %)

Temario

0. Presentación del curso

- Presentación de los participantes.
- Expectativas recíprocas.

1. Análisis de datos con `R`

1.1. El paquete de software estadístico R

El paquete estadístico `R` ofrece un entorno de trabajo y un conjunto de herramientas muy completo para la carga, manejo y análisis de datos cuantitativos. Comparado con otras herramientas como SPSS o Stata `R` tiene algunas ventajas que conviene señalar:

1. Es de fuente abierta (open source) por lo que no tiene costo alguno y no estamos sujetos a las políticas de licencias o precios de una empresa.

2. Cuenta con una comunidad de desarrolladores amplia, por lo que se actualiza permanentemente e incorpora frecuentemente innovaciones en métodos estadísticos.

3. Cuenta con una gran comunidad de usuarios, diversa, amable y solidaria que se encarga de manera voluntaria de documentar el software y proveer soporte técnico a través de sitios de preguntas como StackOverflow.

4. La metodología de trabajo implícita en R favorece la reproducibilidad de resultados e impide la corrupción de datos.

5. Es multiplataforma: funciona en Windows, Mac y Linux.

No obstante es necesario aclarar a los interesados e interesadas en el curso deberán estar dispuestos a aprender un paradigma de uso de software diferente al usual. Si bien algunas operaciones de `R` pueden hacerse seleccionando opciones través de menus, la mayor parte del trabajo se lleva a cabo escribiendo comandos en una consola. En sentido estricto `R` es un lenguaje de programación y un entorno de desarrollo de software aplicado al análisis de datos más que un paquete de análisis estadístico. A pesar de la dificultad que este cambio de paradigma puede significar no hay que perder de vista que `R` es un lenguaje de programación fácil de aprender dentro de su género.

En esta sesión cubriremos:

- Instalación de `R` y RStudio, conexión con CRAN e instalación de librerías adicionales.
- Características básicas de `R`: objetos, tipos y estructuras de datos, funciones y operadores.
- Sintaxis de funciones de uso frecuente en `R`.
- Importación de datos desde Excel, SPSS, Stata, .csv, SAS, etc.

1.2. Manejo y exploración de datos con `R`

Un hecho inevitable del análisis cuantitativo es que pasaremos más tiempo limpiando y ordenando nuestros datos que analizándolos. `R` tiene varias funciones que simplifican este proceso y minimizan el tiempo que le dedicamos. Además nos permite manejar los datos de forma flexible: generar subconjuntos, ordenarlos, filtrarlos y extraer medidas sumarias. Utilizando las funciones de las librerías del metapaquete `tidyverse` aprenderemos a realizar operaciones muy complejas encadenando comandos muy simples. Además de la facilidad de uso, `tidyverse` tiene la ventaja de ser totalmente escalable: con una misma librería

trabajamos con bases de datos de unos kilobytes a unos cientos de gigabytes.

- Reestructuración de datos: el paradigma *tidy data*.[1]
- Codificación y recodificación de variables.
- Filtrado y selección de variables.
- Sumarios de variables.
- Conteos y proporciones.
- Medidas de tendencia central y dispersión.

1.3.1. Gráficos en `R`

Una de las características más potentes de `R` es la capacidad de generar gráficos completamente personalizados. `R` cuenta con varios motores de gráficos: la función básica `plot(x)` produce, con comandos muy simples, gráficos básicos para exploración, mientras que `ggplot` nos permite personalizar completamente la geometría y apariencia de los gráficos. Con gran poder viene gran responsabilidad: el estilo minimalista de Cleveland y Tufte nos servirá de guía para hacer gráficos tan elocuentes como rigurosos y legibles, evitando información superflua (chartjunk) y un manejo engañoso de las escalas.

- Como graficar datos. Teoría y uso a partir de William Cleveland y Edward Tufte.
- Elementos básicos de un gráfico de datos.
- Tipos básicos de gráficos (Dispersión, puntos, caja, kernel de densidad)

1.3.2. Gráficos en `R` (II)

Personalización de gráficos en `R` con la librería `ggplot2`.

- Anotaciones y etiquetado condicional.
- Combinación elementos geométricos en un gráfico.
- Gráficos por paneles.

1.4. Aplicación

Sesión práctica de análisis descriptivo de datos. Se sugiere tener identificada, para esta sesión, una base de datos de interés sobre la que realizar el ejercicio.

[1] Una columna es una variable y una fila una observación.

Análisis de datos

2.1. Sumarios descriptivos

Uno de los usos más prácticos de la estadística descriptiva es la obtención de sumarios: medidas diseñadas para presentar de manera resumida la información contenida en una variable o la relación entre dos o más variables.

- Medidas de tendencia central
  - Media aritmética
  - Mediana de orden
  - Moda
- Medidas de dispersión
  - Varianza y desviación estándar
  - Rango intercuartil

2.2. Tablas de contingencia

Las tablas de contingencia fueron el método pionero para datos categóricos. Más allá de su utilidad intrínseca también nos sirven para introducir dos conceptos fundamentales del análisis cuantitativo en el marco frecuentista: las pruebas de hipótesis y los modelos de independencia.

- Creación y uso de tablas de contingencia.
  - Tablas simples, tablas con conteos marginales, tablas con proporciones.
  - Tablas de más de dos dimensiones.
  - Uso de las funciones `table()` y `count()` para bases de datos ponderadas.

2.3. Modelos de independencia.

- Modelo de independencia estadística χ^2 para tablas contingencia de dos dimensiones.
  - Análisis de los residuos de χ^2 y partición de χ^2.
  - Métodos gráficos para el análisis de tablas de contingencia con el paquete `vcd`.
- ¿Qué es -y qué *no* es- una prueba de hipótesis?[2]
  - Pruebas de hipótesis en el marco frecuentista.
  - Más allá del ritual de los *p-value*.

[2] O por qué, lamentablemente, $p(H_0|D) \neq p(D|H_0)$.

2.4. Medidas de asociación entre variables categóricas y continuas.

Estimar la asociación de dos variables es un paso fundamental en el análisis de datos, nos permite pasar de describir variables a establecer relaciones entre ellas. En esta unidad consideraremos especialmente tipos de correlación para datos categóricos, nominales u ordinales y funciones para generar matrices de covarianza heterogéneas. La introducción de matrices de covarianza servirá también como introducción al análisis factorial.

- Medidas de asociación para variables continuas.
- Coeficiente ϕ, V de Cramer.
- Coeficientes de correlación policórica y tetracórica.
- Matrices de correlación y covarianza.
- Gráficos de matriz de correlación y de red de correlaciones.

Modelos lineales generalizados

3.1. Modelos lineales

Los modelos lineales nos permiten modelar numéricamente hipótesis explicativas: expresar el valor de una variable como función de otra u otras. Son una de las herramientas más aplicadas del análisis cuantitativo en las ciencias sociales, y ello por buenos motivos. Con frecuencia los fenómenos que investigamos incluyen múltiples factores explicativos y es necesario distinguir a los relevantes de los irrelevantes, así como cuantificar la importancia de cada uno cuando todos los demás están presentes. Las cantidades de interés que extraemos del ajuste de un modelo lineal nos permiten hacer estas operaciones.

- Ajuste de modelos lineales con `R` usando la función `lm()`
    - Interpretación de los coeficientes (pendientes), errores estándar y pruebas de significacia.
    - Estadísticos de bondad de ajuste y pruebas de análisis de varianza (ANOVA) para dos modelos.
    - Supuestos de los modelos lineales.

3.2. Modelos lineales para variables dependientes dicotómicas (GLM)

Aunque tienen su origen en las variables continuas es posible extenderlos a variables dependientes categóricas, dicotómicas o politómicas, a partir de distribuciones binomiales y polinomiales. En esta unidad conoceremos los aspectos básicos de los modelos lineales, enfatizando sobre todo su aplicación a hipótesis de las ciencias sociales.

- Modelos logit binomiales.
  - Ajuste de modelos con la función `glm()`
  - Interpretación de los coeficientes, errores estándar y significancia.
  - Estadísticos de bondad de ajuste, prueba ANOVA para dos modelos.

3.3. Modelos lineales para variables dependientes politómicas.

- Modelos logit polinomiales.
  - Ajuste de modelos logit polinomiales con `glm()`.
  - Interpretación de coeficientes.
  - Diagnóstico.

3.4. Aplicación.

Aplicación de modelos lineales para controlar los efectos de distintas variables en las trayectorias escolares.

- Blanco, Solís, Robles, (2015) Caminos desiguales. Trayectorias educativas y laborales de los jóvenes en la Ciudad de México.

Bibliografía

Agresti, Alan. 2002. *Categorical data analysis*. 2nd ed. Wiley series in probability and statistics. New York: Wiley-Interscience.

Baranger, Denis. 2012. *Epistemología y metodología en la obra de Pierre Bourdieu*. Segunda. Posadas.

Byrne, D. S., y Charles C. Ragin, eds. 2009. *The SAGE handbook of case-based methods*. Los Angeles ; London: SAGE.

Finch, W. Holmes, y Brian F. French. 2015. *Latent variable modeling with R*. New York: Routledge, Taylor & Francis Group.

Greenacre, Michael J, y Fundación BBVA. 2008. *La Práctica del análisis de correspondencias*. Barcelona: Fundación BBVA.

Husson, François, Sébastien Lê, y Jérôme Pagès. 2011. *Exploratory multivariate analysis by example using R*. Chapman & Hall/CRC computer science and data analysis. Boca Raton: CRC Press.

Husson, François, Lê, Sébastien, y Pagès, Jérôme. 2013. *Análisis de datos con R*. Bogotá: Escuela Colombiana de Ingeniería.

Kabacoff, Robert. 2011. *R in action: data analysis and graphics with R*. Shelter Island, NY: Manning.

Kaplan, David, y Sage Publications, eds. 2004. *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, Calif: Sage.

Le Roux, Brigitte, y Henry Rouanet. 2010. *Multiple correspondence analysis*. Quantitative applications in the social sciences 163. Thousand Oaks, Calif: Sage Publications.

Robertson, Judy, y Maurits Kaptein, eds. 2016. *Modern Statistical Methods for HCI*. Human–Computer Interaction Series. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-26633-6.

Véliz, Carlos. 2016. *Análisis Multivariante. Métodos Estadísticos Multivariantes para la Investigación*. Ciudad de México: Cengage.