

SOC3070 ANÁLISIS DE DATOS CATEGÓRICOS

Segundo Semestre 2021

Instructor:	Mauricio Bucca	Horario Clase:	Lunes 10:00 - 12:50
Email:	mebucca@uc.cl	Horario Consulta:	Viernes 9:00 - 10:00

Ayudantía:

- Ayudante: Roberto Velázquez (ravelasq@uc.cl)
- Horario: Miércoles 14:00 - 16:50

Páginas del curso:

- Canvas
- Repositorio Github: https://github.com/mebucca/cda_soc3070

Descripción: Este curso se centra en el estudio de métodos estadísticos para el análisis de datos categóricos, con especial énfasis en modelos de regresión para variables dependientes categóricas. Este tipo de modelos es particularmente importante dado que muchos fenómenos de interés para las ciencias sociales no se manifiestan de modo continuo sino binario (si/no), ordinal (alto/medio/bajo), nominal (candidatos A/B/C) o como recuento (números de contagios por covid-19). Cuando trabajamos con variables dependientes categóricas los modelos estándar de regresión lineal son, *por lo general*, inapropiados. En consecuencia, el curso cubrirá una amplia gama de métodos disponibles para modelar este tipo de datos, muchos de los cuales son instancias especiales de los así llamados *Modelos Lineales Generalizados* (GLM). La clase tiene como objetivo desarrollar los conocimientos y habilidades necesarias para que los estudiantes apliquen estos métodos en su propia investigación empírica.

Resultados esperados: Al completar este curso *satisfactoriamente* se espera que los estudiantes sean capaces de:

- Entender los fundamentos estadísticos y matemáticos de los principales modelos para el análisis de datos categóricos.
- Implementar estos métodos en el software estadístico R (o Stata).
- Interpretar sustantivamente los resultados de estos métodos.
- Manipular matemáticamente los resultados de estos métodos para producir "cantidades de interés".

Prerequisitos sugeridos:

- Un curso de introducción a la estadística de nivel pregrado.
- Un curso de introducción al análisis de regresión de nivel pregrado.
- Familiaridad con el software estadístico R (o Stata)

Software: Las clases y evaluaciones del curso involucrarán el uso de software estadístico. El contenido expositivo del curso y ayudantías se desarrollará exclusivamente en R. No obstante lo anterior, los estudiantes tienen libertad para desarrollar las evaluaciones tanto en R como Stata. El código de R que acompaña cada clase será puesto a disposición de los estudiantes.

Contenidos:

1. Fundamentos de teoría de la Probabilidad
2. Variables Aleatorias y Distribuciones Discretas
3. Estimación via Maximum Likelihood (MLE)
4. Tablas de Contingencia
5. Modelo Lineal de Probabilidad (LPM)
6. Modelos Lineales Generalizados (GLM)
7. Modelos de Regresión Logística
8. Modelos de Regresión Logística Ordenada
9. Modelos de Regresión Logística Multinomial
10. Modelos de Regresión Poisson y Quasi-Poisson

Evaluaciones: Tareas cortas (30%), Trabajos (40%), Trabajo final (30%).

- Tareas cortas: 5 tareas cortas con una ponderación de 6% cada una.
- Trabajos: 2 trabajos de mediana extensión. Cada trabajo tiene una ponderación de 20%
- Trabajo final: 1 trabajo final con ponderación de 30% consistente en replicar resultados seleccionados de un artículo publicado en una revista científica. El artículo será elegido por el profesor y los datos serán provistos a los estudiantes. Los estudiantes deberán replicar los resultados, reportarlos en formato de informe y entregar un código de replicación.

Es decir, la nota final de cada estudiante i está dada por la suma ponderada de sus notas en las tareas cortas (TC_j), trabajos (T_j) y el trabajo final (TF):

$$NF_i = \sum_{j=1}^5 TC_{ij} \times 0.06 + \sum_{j=1}^2 T_{ij} \times 0.2 + TF_i \times 0.3$$

Bibliografía: El curso es auto-contenido y no considera un texto obligatorio. No obstante, se recomiendan los siguientes libros como textos de consulta.

- Agresti, Alan. (1996). Introduction to Categorical data Analysis. Second Edition. Wiley Series.
- Gelman, A. y Hill, J. (2007). Data Analysis Using Regression and Multilevel/ Hierarchical Models. Cambridge University Press.
- Bilder, Christopher R., and Thomas M. Loughin (2014). Analysis of categorical data with R. CRC Press [Disponible online por medio de Biblioteca UC].

Políticas del curso:

- **Asistencia:** La asistencia regular a clases es importante, esperada, pero no obligatoria.
- **Entregas:** Las tareas y trabajos deben ser entregados a más tardar el día indicado en el calendario del curso a más tardar a las 23:59pm.
- **Email:** Escribir primariamente para agendar reunión en el horario de consulta. Incluir SOC3070 en el asunto del correo. Respondo correos una vez al día (en la mañana), de lunes a viernes.
- **Honestidad Académica:** El desconocimiento de la política de honestidad académica de la universidad no es una explicación razonable para su incumplimiento.