Maestría en Estadística Aplicada
Universidad Nacional de Córdoba
Cohorte 2020 -2021

Proyecto de Análisis de Datos I
Docentes a cargo: Dra. María Gabriela Palacio – Mgtr. Sergio Martín Buzzi

---

Guía Nº 3

El objetivo de Proyecto de Análisis de Datos I es adquirir destrezas en el análisis estadístico de datos. Se presentarán en cada encuentro un conjunto de problemas con un cuestionario a responder. Esta tercera y última guía consiste en dos ejercicios, uno de modelos log lineales y otro sobre análisis de cluster. La resolución debe realizarse en grupos de hasta 4 personas en dos archivos separados, preferentemente en formato pdf (uno por cada problema, con nombre: Problema1ApellidoApellidoApellidoApellido.pdf) y luego subirlos al aula virtual de la Maestría en Estadística Aplicada (sólo uno de los integrantes del grupo). Pueden usar el programa estadístico que deseen (InfoStat, R, etc.).

Fecha de entrega del trabajo (en el aula virtual): **jueves 10 de junio de 2021**.

Por cualquier eventualidad, pueden escribirnos a nuestros correos electrónicos: María Gabriela Palacio (gpalacio@exa.unrc.edu.ar) y Sergio Martín Buzzi (sergio.buzzi@unc.edu.ar).

Problema 1

Trabajaremos con la misma base de datos del Problema 1 de la Guía 2.

El archivo *"muestra 80 datos.xlsx"* es la submuestra con la cual seguiremos trabajando.

Considere que el objetivo es investigar asociación entre variables cualitativas, usando Modelos log lineales.

A) Para la Situación 1: dos variables binarias.

   i) Escribe brevemente el objetivo que plantea usando el modelo indicado.

   ii) Realiza un análisis descriptivo de la información que incluya **tablas, gráficos y sus interpretaciones relacionadas al problema**.

   iii) Usando un software y **mostrando la salida del mismo**

1. Escribir el modelo estimado e **interpretar para el problema los coeficientes**.
2. Probar la significación de los parámetros del modelo.
3. Calcula el Odds Ratio (OR) e interpretarlo.
4. ¿A qué conclusión llega?

B) Para la Situación 2: tres variables binarias (las dos de la Situación 1 y una más).

i) Escribe brevemente el objetivo que plantea usando el modelo indicado.

ii) Realiza un análisis descriptivo de la información que incluya **tablas, gráficos y sus interpretaciones relacionadas al problema**.

iii) Escribe el modelo y las hipótesis a probar (**estadísticamente y en palabras en función del problema**).

iv) Usando un software y **mostrando la salida del mismo**

1- Escribir el modelo estimado e **interpretar para el problema los coeficientes** (los no explicados anteriormente).

2- Probar la significación de los parámetros del modelo **explicando lo que ésta significa para el problema**.

3- Ajustar diferentes modelos (desde el más complejo al más simple) analizando la bondad del ajuste con el AIC. ¿Cuál de los modelos resulta mejor?

4- ¿A qué conclusión llega?

Problema 2

Este problema emplea una versión reducida de la base de datos usada en el Problema 2 de la Guía II de la materia, con la finalidad de realizar un análisis de clusters. En este caso solo se emplean los barrios que tienen 5000 o mas habitantes para simplificar la actividad.

En el archivo “proporciones_reducida.xlsx”, se encuentran las siguientes proporciones:

- Proporción de los jefes o jefas de hogar según nivel educativo: inicial (incluye sin instrucción), priminc, primcomp, secinc, seccomp, terinc, tercomp, univinc, univcomp, posinc, poscomp.
- Proporción de las personas de 3 años o más, según si saben leer y escribir: sabeleer, nosabeleer.
- Proporción de personas según cobertura de salud / obra social / prepaga: cobertura, nocobertura.
- Proporción de personas de 14 años o más según condición de ocupación: ocupados, desocupados, inactivos.
- Proporción de personas por tipo de necesidades básicas insatisfechas: nbi1, nbi2, nbi3, nbi4, nbi5.
- Proporción de personas según índice de privación material en el hogar: sin pro, privpat, privrec, privconv.
- Proporción de hogares por tipo: casa, rancho, casilla, departamento, piezaolocal.
- Proporción de hogares con determinada cantidad de habitantes: uno, dos, tres, cuatro, cinco, seis, 7, ochoomas.

A partir de dichos datos, se pide:

A) Aplicar al menos un método **jerárquico**, por ejemplo: vecino mas cercano, vecino mas lejano, vinculación promedio o Ward.

   i) Describir muy brevemente en qué consiste.

   ii) Graficar el dendograma.

   iii) Determinar cuántos grupos se seleccionan. Justificar, por ejemplo usando el Criterio de Hartigan o el Coeficiente de Correlación Cofenético.

B) Aplicar el método **k-means**

   i) Indicar cómo se puede atemperar la influencia de la elección de los centroides iniciales en los resultados.

   ii) Determinar si se encontró una estructura de agrupación fuerte o débil; por ejemplo, empleando el Coeficiente de Silhouette.

C) En base a alguno de los agrupamientos previos, realizar un análisis descriptivo comparativo **entre grupos**.

   i) En base al mismo, caracterice a los grupos.

   ii) ¿Se le pueden dar "nombres" intuitivos a los grupos?

D) Indicar si se encuentra alguna conexión entre los resultados obtenidos en los incisos anteriores y los obtenidos en el Problema 2 de la guía II.

   i) Opcional: realizar un análisis factorial exploratorio breve usando esta base reducida para que sean más comparables los resultados.