Maestría en Estadística Aplicada
Universidad Nacional de Córdoba
Cohorte 2022 -2023

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

---

Guía Nº 1

Esta primera guía consiste en dos ejercicios, uno de análisis de la varianza y otro de análisis de regresión.

La resolución debe realizarse en grupos de hasta 4 personas.

Para cada uno de los dos problemas de esta guía se debe subir al aula virtual de la Maestría un archivo .pdf con un desarrollo completo y un archivo .R, .Rmd, o .ipynb documentado.

Los archivos deberán ser subidos por un solo integrante de cada grupo y los nombres de estos deberán seguir el formato: ProblemaXApellidoApellidoApellidoApellido.

Fecha de entrega: **lunes 14 de abril de 2025**.

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

El paquete MASS de R contiene la base de datos UScereal que tiene 65 filas y 11 columnas. Los datos provienen de la Exposición de Gráficos Estadísticas de la ASA de 1993 y se toman de la etiqueta alimentaria obligatoria de la F&DA correspondiente a cajas de cereales comerciales que se venden en supermercados.

a) Describa los datos, indicando cuáles son las variables y de qué tipo son, lo que representa cada observación y cómo considera que se realizó la recolección de datos.
b) Para realizar un Análisis de la varianza (ANOVA), ¿qué variables del archivo de datos tendría sentido seleccionar como variables respuestas y como factores? Explicar brevemente.
c) Considere como factor al Fabricante.
    a) Explique qué significa considerar a Fabricante como “factor fijo” y que implica respecto a las conclusiones del análisis.

b) Seleccionar 1 (una) variable respuesta y para ella:

i) Indicar cuál sería el objetivo al que responde el ANOVA en este caso.

ii) ¿Considera adecuado realizar este análisis dada la manera en que se recolectan los datos? Justifique su respuesta.

iii) Escriba el modelo lineal, indicado lo que significa cada uno de sus términos.

iv) Para los términos asociados a variables aleatorias indique cuáles son los supuestos subyacentes.

v) Plantee las hipótesis a probar, en términos de los parámetros del modelo.

vi) Complete la siguiente Tabla ANOVA consignando un nombre apropiado para la Fuente de Variación y los grados de libertad correspondientes a cada fuente de variación:

Fuentes de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio
……..			
Error			
Total			

vii) Si en la Tabla anterior SC1 es la primera Suma de cuadrados, y SC2 es la segunda, ¿cómo se obtiene el valor del Estadístico de la prueba y cuál es su distribución (incluyendo los grados de libertad bajo hipótesis nula).

viii) A partir de la información provista en la base de datos, y usando software:

- Realice el análisis descriptivo adecuado (según el objetivo descrito en i) y comente lo que observa. (NOTA: Incluya diagramas de caja y comentarios de lo que sugieren respecto de la situación, si se presentan valores extremos, si cree que se cumplirán los supuestos).
- Indique las conclusiones que se desprenden de la tabla ANOVA.
- Determine (gráficamente y con pruebas de hipótesis) si las conclusiones del análisis resultan válidas, explicando los análisis que realiza. ¿Qué variable se

usa para realizar estos análisis? ¿qué ventaja tiene usar esta variable?

- En caso necesario realiza una transformación para resolver al menos uno de los problemas detectados, muestre los análisis realizados y comente lo que encuentra. Mencione brevemente otras alternativas para proceder en lugar de realizar transformación.
- Realiza, de ser necesario, un test de comparaciones múltiples apropiado, indicando la información que brinda respecto a la situación.

IV) Indique otra variable de la base de datos que podría considerarse como Factor para un ANOVA y describa cuál sería en este caso el objetivo del análisis.

Problema 2

Los datos del archivo base_regresion.xls corresponden a 5396 alojamientos alquilados por medio de la plataforma Airbnb en la Ciudad de Buenos Aires durante el día 25/11/2015[1]. La base citada contiene información sobre el identificador del alojamiento (id), el precio de una noche de alojamiento medido en dólares (precio), el tipo de alojamiento (tipo), la cantidad de valoraciones que ha recibido el alojamiento (valoraciones), el puntaje promedio de un máximo de 5 que recibió el alojamiento (puntaje), la cantidad de personas que puede recibir el alojamiento (personas), la cantidad de dormitorios (dormitorios), la cantidad de baños (baños), la estadía mínima (estadia), la distancia a la estación de trenes mas cercana (distancia) y la cantidad de dependencias culturales que se encuentran a menos de 300 metros a la redonda (dependencias).

El objetivo general del problema es ajustar un modelo de regresión para explicar precio en dólares pagado por la estadía por una noche (precio).

a) Realice un análisis descriptivo de las variables utilizando medidas resumen y gráficos.

b) De acuerdo con lo observado en el análisis descriptivo de la variable precio, ¿considera conveniente transformarla tomando logaritmo natural para especificar el modelo de regresión?

[1] Los datos son provenientes de la página de Tom Slee (http://tomslee.net/airbnb-data).

c) Estime un modelo de regresión lineal sin considerar las variables tipo, distancia y dependencias.

   1) ¿Son significativas todas las variables explicativas incluidas?
   2) Interprete los coeficientes estimados.
   3) ¿El modelo tiene buen poder explicativo? ¿Cómo lo mide?
   4) ¿Se cumplen los supuestos en los que se basa el modelo? Justifique adecuadamente.

d) Mejore la especificación del modelo agregando o quitando variables, sin considerar la variable tipo.

   1) ¿Qué criterio utilizó para la selección de variables?
   2) Verifique el cumplimiento de los supuestos.
   3) Compare el poder explicativo del nuevo modelo con el estimado en el inciso c. Justifique adecuadamente.
   4) Interprete los coeficientes del nuevo modelo.

e) Suponga que se desea estudiar el efecto que tienen los distintos tipos de alojamientos (tipo) en el precio, ¿Cómo podemos incorporar dicho análisis en el modelo de regresión?

   1) Estime el nuevo modelo.
   2) Interprete los coeficientes estimados.
   3) Evalúe el poder explicativo del modelo que seleccionó. Compare con los modelos anteriores.
   4) ¿Se verifican los supuestos del modelo? Justifique adecuadamente.

5) Existen valores atípicos y/o valores influyentes? En caso afirmativo explique cómo los detecta y que decisión toma respecto a ellos para continuar con el análisis estadístico.